(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
2019 年 1 月 24 日 (24.01.2019)

WIPO | PCT

(10) 国际公布号
WO 2019/015641 A1

(51) 国际专利分类号:
G06K 9/00 (2006.01) *G06K 9/62* (2006.01)

(21) 国际申请号: PCT/CN2018/096252

(22) 国际申请日: 2018 年 7 月 19 日 (19.07.2018)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
201710592780.7 2017年7月19日 (19.07.2017) 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, Grand Cayman (KY)。

(72) 发明人; 及

(71) 申请人 (仅对*US*): 江南(JIANG, Nan) [CN/CN]; 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)。

赵宏伟(ZHAO, Hongwei) [CN/CN]; 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: 北京国昊天诚知识产权代理有限公司(CO-HORIZON INTELLECTUAL PROPERTY INC.); 中国北京市朝阳区小关北里甲2号渔阳置业大厦B座605, Beijing 100029 (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, 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,

(54) **Title:** MODEL TRAINING METHOD AND METHOD, APPARATUS, AND DEVICE FOR DETERMINING DATA SIMILARITY

(54) **发明名称**: 模型的训练方法、数据相似度的确定方法、装置及设备

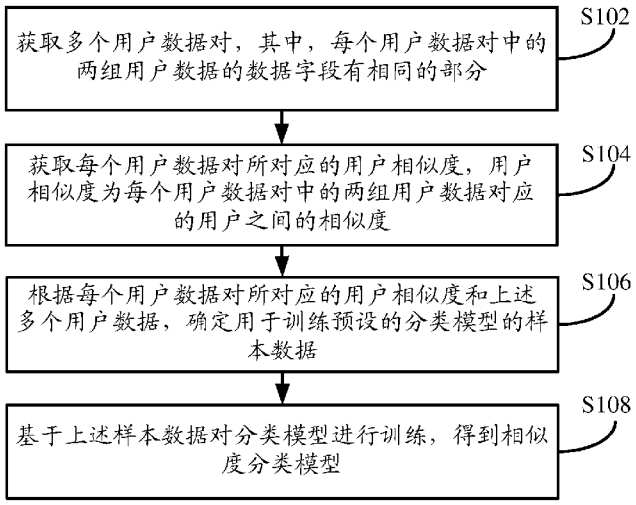

图 1

(57) **Abstract:** Disclosed in the embodiments of the present application are a model training method and a method, an apparatus, and a device for determining data similarity, the model training method comprising: acquiring a plurality of user data pairs, the data fields of the two sets of user data in each user data pair having an identical part; acquiring the user similarity corresponding to each user data pair, the user similarity being the similarity between the users corresponding to the two sets of user data in each user data pair; on the basis of the user similarity corresponding to each user data pair and the plurality of user data pairs, determining sample data used for training a preset classification model; and, on the basis of the sample data, training the classification model to obtain a similarity classification model. The embodiments of the present application can implement rapid model training, increasing model training efficiency and reducing resource consumption.

WO 2019/015641 A1

[见续页]

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))。

(57) 摘要：本申请实施例公开了一种模型的训练方法、数据相似度的确定方法、装置及设备，该模型的训练方法包括：获取多个用户数据对，其中，所述每个用户数据对中的两组用户数据的数据字段有相同的部分；获取每个用户数据对所对应的用户相似度，所述用户相似度为每个用户数据对中的两组用户数据对应的用户之间的相似度；根据所述每个用户数据对所对应的用户相似度和所述多个用户数据对，确定用于训练预设的分类模型的样本数据；基于所述样本数据对所述分类模型进行训练，得到相似度分类模型。利用本申请实施例，可以实现模型的快速训练，提高模型训练效率并减少资源消耗。

模型的训练方法、数据相似度的确定方法、装置及设备

技术领域

本申请涉及计算机技术领域，尤其涉及一种模型的训练方法、数据相似度的确定方法、装置及设备。

背景技术

人脸识别作为一种新型的身份核实方式，在为用户提供便利的同时也产生了新的风险点。对于长相极为相似的多个用户（如双胞胎），通过人脸识别将很难有效区分不同用户，从而极易造成因为无法正确识别导致的账户误登录，以及账户资金被盗用等风险。双胞胎特别是同卵双胞胎作为已知的长相极为相似的最典型情况，因为两者彼此关系亲密，非常容易产生上述风险行为。如何从大量数据中确定双胞胎的用户数据成为需要解决的重要问题。

通常，基于监督式的机器学习方法利用预先选取的样本数据构造识别模型，具体地，调查人员通过问卷调查、有奖问答或人工观察等方式进行社会调查，收集用户数据，并通过人工观察或向调查者询问等方式得到的用户之间的关联关系或双胞胎关系进行标注。通过人工标注的关联关系或双胞胎关系，使用相应的用户数据作为样本数据构造识别模型。

然而，上述通过监督式机器学习方法构造的识别模型，其样本数据需要进行人工标注，而人工标注的过程会消耗大量的人力资源，而且还会消耗大量的时间进行标注，从而使得模型训练效率低下，且资源消耗较大。

发明内容

本申请实施例的目的是提供一种模型的训练方法、数据相似度的确定方法、装置及设备，以实现模型训练的快速完成，提高模型训练效率并减少资源消耗。

为解决上述技术问题，本申请实施例是这样实现的：

本申请实施例提供的一种模型的训练方法，所述方法包括：

获取多个用户数据对，其中，所述每个用户数据对中的两组用户数据的数据字段有相同的部分；

获取每个用户数据对所对应的用户相似度，所述用户相似度为每个用户数

据对中的两组用户数据对应的用户之间的相似度;

根据所述每个用户数据对所对应的用户相似度和所述多个用户数据对，确定用于训练预设的分类模型的样本数据;

基于所述样本数据对所述分类模型进行训练，得到相似度分类模型。

可选地，所述获取每个用户数据对所对应的用户相似度，包括:

获取第一用户数据对所对应的用户的生物特征，其中，所述第一用户数据对为所述多个用户数据对中的任意用户数据对;

根据所述第一用户数据对所对应的用户的生物特征，确定所述第一用户数据对所对应的用户相似度。

可选地，所述生物特征包括面部图像特征，

所述获取第一用户数据对所对应的用户的生物特征，包括:

获取第一用户数据对所对应的用户的面部图像;

对所述面部图像进行特征提取，得到面部图像特征;

相应的，所述根据所述第一用户数据对所对应的用户的生物特征，确定所述第一用户数据对所对应的用户相似度，包括:

根据所述第一用户数据对所对应的用户的面部图像特征，确定所述第一用户数据对所对应的用户相似度。

可选地，所述生物特征包括语音特征，

所述获取第一用户数据对所对应的用户的生物特征，包括:

获取第一用户数据对所对应的用户的语音数据;

对所述语音数据进行特征提取，得到语音特征;

相应的，所述根据所述第一用户数据对所对应的用户的生物特征，确定所述第一用户数据对所对应的用户相似度，包括:

根据所述第一用户数据对所对应的用户的语音特征，确定所述第一用户数据对所对应的用户相似度。

可选地，所述根据所述每个用户数据对所对应的用户相似度和所述多个用户数据对，确定用于训练分类模型的样本数据，包括:

对所述多个用户数据对中的每个用户数据对进行特征提取，得到每个用户数据对中的两组用户数据之间相关联的用户特征;

根据所述每个用户数据对中用户数据之间相关联的用户特征和所述每个用户数据对所对应的用户相似度，确定用于训练分类模型的样本数据。

可选地，所述根据所述每个用户数据对中的两组用户数据之间相关联的用户特征和所述每个用户数据对所对应的用户相似度，确定用于训练分类模型的样本数据，包括:

根据每个用户数据对所对应的用户相似度和预定的相似度阈值，从所述多个用户数据对所对应的用户特征中选取正样本特征和负样本特征;

将所述正样本特征和负样本特征作为用于训练分类模型的样本数据。

可选地，所述用户特征包括户籍维度特征、姓名维度特征、社交特征和兴趣爱好特征；所述户籍维度特征包括用户身份信息的特征，所述姓名维度特征包括用户姓名信息的特征和用户姓氏的稀缺程度的特征，所述社交特征包括用户的社会关系信息的特征。

可选地，所述正样本特征和负样本特征中包含的特征数目相同。

可选地，所述相似度分类模型为二分类器模型。

本申请实施例还提供的一种数据相似度的确定方法，所述方法包括:

获取待测用户数据对;

对所述待测用户数据对中每组待测用户数据进行特征提取，得到待测用户特征;

根据所述待测用户特征和预先训练的相似度分类模型，确定所述待测用户数据对中的两组待测用户数据对应的用户之间的相似度。

可选地，所述方法还包括:

如果所述待测用户数据对中的两组待测用户数据对应的用户之间的相似度大于预定相似度分类阈值，则确定所述待测用户数据对所对应的待测用户为双胞胎。

本申请实施例提供的一种模型的训练装置，所述装置包括:

数据获取模块，用于获取多个用户数据对，其中，所述每个用户数据对中的两组用户数据的数据字段有相同的部分;

相似度获取模块，用于获取每个用户数据对所对应的用户相似度，所述用户相似度为每个用户数据对中的两组用户数据对应的用户之间的相似度;

样本数据确定模块，用于根据所述每个用户数据对所对应的用户相似度和所述多个用户数据对，确定用于训练预设的分类模型的样本数据;

模型训练模块，用于基于所述样本数据对所述分类模型进行训练，得到相似度分类模型。

可选地，所述相似度获取模块，包括:

生物特征获取单元，用于获取第一用户数据对所对应的用户的生物特征，其中，所述第一用户数据对为所述多个用户数据对中的任意用户数据对;

相似度获取单元，用于根据所述第一用户数据对所对应的用户的生物特征，确定所述第一用户数据对所对应的用户相似度。

可选地，所述生物特征包括面部图像特征，

所述生物特征获取单元，用于获取第一用户数据对所对应的用户的面部图像；对所述面部图像进行特征提取，得到面部图像特征;

相应的，所述相似度获取单元，用于根据所述第一用户数据对所对应的用户的面部图像特征，确定所述第一用户数据对所对应的用户相似度。

可选地，所述生物特征包括语音特征，

所述生物特征获取单元，用于获取第一用户数据对所对应的用户的语音数据；对所述语音数据进行特征提取，得到语音特征;

相应的，所述相似度获取单元，用于根据所述第一用户数据对所对应的用户的语音特征，确定所述第一用户数据对所对应的用户相似度。

可选地，所述样本数据确定模块，包括:

特征提取单元，用于对所述多个用户数据对中的每个用户数据对进行特征提取，得到每个用户数据对中的两组用户数据之间相关联的用户特征;

样本数据确定单元，用于根据所述每个用户数据对中的两组用户数据之间相关联的用户特征和所述每个用户数据对所对应的用户相似度，确定用于训练分类模型的样本数据。

可选地，所述样本数据确定单元，用于根据每个用户数据对所对应的用户相似度和预定的相似度阈值，从所述多个用户数据对所对应的用户特征中选取正样本特征和负样本特征；将所述正样本特征和负样本特征作为用于训练分类模型的样本数据。

可选地，所述用户特征包括户籍维度特征、姓名维度特征、社交特征和兴趣爱好特征；所述户籍维度特征包括用户身份信息的特征，所述姓名维度特征包括用户姓名信息的特征和用户姓氏的稀缺程度的特征，所述社交特征包括用户的社会关系信息的特征。

可选地，所述正样本特征和负样本特征中包含的特征数目相同。

可选地，所述相似度分类模型为二分类器模型。

本申请实施例还提供的一种数据相似度的确定装置，所述装置包括：

待测数据获取模块，用于获取待测用户数据对；

特征提取模块，用于对所述待测用户数据对中每组待测用户数据进行特征提取，得到待测用户特征；

相似度确定模块，用于根据所述待测用户特征和预先训练的相似度分类模型，确定所述待测用户数据对中的两组待测用户数据对应的用户之间的相似度。

可选地，所述装置还包括：

相似度分类模块，用于如果所述待测用户数据对中的两组待测用户数据对应的用户之间的相似度大于预定相似度分类阈值，则确定所述待测用户数据对所对应的待测用户为双胞胎。

本申请实施例提供的一种模型的训练设备，所述设备包括：

处理器；以及

被安排成存储计算机可执行指令的存储器，所述可执行指令在被执行时使所述处理器执行以下操作：

获取多个用户数据对，其中，所述每个用户数据对中的两组用户数据的数据字段有相同的部分；

获取每个用户数据对所对应的用户相似度，所述用户相似度为每个用户数据对中的两组用户数据对应的用户之间的相似度；

根据所述每个用户数据对所对应的用户相似度和所述多个用户数据对，确定用于训练预设的分类模型的样本数据；

基于所述样本数据对所述分类模型进行训练，得到相似度分类模型。

本申请实施例提供的一种数据相似度的确定设备，所述设备包括：

处理器；以及

被安排成存储计算机可执行指令的存储器，所述可执行指令在被执行时使所述处理器执行以下操作：

获取待测用户数据对；

对所述待测用户数据对中每组待测用户数据进行特征提取，得到待测用户特征；

根据所述待测用户特征和预先训练的相似度分类模型，确定所述待测用户数据对中的两组待测用户数据对应的用户之间的相似度。

由以上本申请实施例提供的技术方案可见，本申请实施例通过获取的多个

用户数据对，且每个用户数据对中的两组用户数据的数据字段有相同的部分，以及获取的每个用户数据对所对应的用户相似度，确定用于训练预设的分类模型的样本数据，然后，基于样本数据对分类模型进行训练，得到相似度分类模型，以便后续可以通过相似度分类模型确定待测用户数据对中的两组待测用户数据对应的用户之间的相似度，这样，仅通过相同的数据字段得到多个用户数据对，并通过用户相似度确定每个用户数据对中的两组用户数据对应的用户之间的关联关系，得到用于训练预设的分类模型的样本数据，而不需要人工标注即可得到样本数据，可以实现模型训练的快速完成，提高了模型训练效率并减少资源消耗。

附图说明

为了更清楚地说明本申请实施例或现有技术中的技术方案，下面将对实施例或现有技术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申请中记载的一些实施例，对于本领域普通技术人员来讲，在不付出创造性劳动性的前提下，还可以根据这些附图获得其他的附图。

图 1 为本申请一种模型的训练方法实施例;

图 2 为本申请一种数据相似度的确定方法实施例;

图 3 为本申请一种检测应用程序的界面示意图;

图 4 为本申请一种数据相似度的确定方法实施例;

图 5 为本申请一种数据相似度的确定过程的处理逻辑示意图;

图 6 为本申请一种模型的训练装置实施例;

图 7 为本申请一种数据相似度的确定装置实施例;

图 8 为本申请一种模型的训练设备实施例;

图 9 为本申请一种数据相似度的确定设备实施例。

具体实施方式

本申请实施例提供一种模型的训练方法、数据相似度的确定方法、装置及设备。

为了使本技术领域的人员更好地理解本申请中的技术方案，下面将结合本申请实施例中的附图，对本申请实施例中的技术方案进行清楚、完整地描述，显然，所描述的实施例仅仅是本申请一部分实施例，而不是全部的实施例。基

于本申请中的实施例，本领域普通技术人员在没有作出创造性劳动前提下所获得的所有其他实施例，都应当属于本申请保护的范围。

实施例一

如图 1 所示，本申请实施例提供一种模型的训练方法，该方法的执行主体可以为终端设备或服务器，其中的终端设备可以是个人计算机等，服务器可以是独立的一个服务器，也可以是由多个服务器组成的服务器集群。本申请实施例中为了提高模型训练的效率，该方法的执行主体以服务器为例进行详细说明。该方法具体可以包括以下步骤:

在步骤 S102 中，获取多个用户数据对，其中，每个用户数据对中的两组用户数据的数据字段有相同的部分。

其中，每个用户数据对中可以包含多个不同用户的用户数据，例如，多个用户数据对中包括用户数据对 A 和用户数据对 B，其中，用户数据对 A 中包括用户数据 1 和用户数据 2,用户数据对 B 中包括用户数据 3 和用户数据 4 等。用户数据可以是与某用户相关的数据，例如，用户的姓名、年龄、身高、住址、身份证号码、社会保障卡（即社保卡）号码等身份信息，还可以包括用户的兴趣爱好、购买商品、旅游等信息。数据字段可以是能够表征用户数据对中的两组不同用户数据对应的用户的身份，以及用户之间的关联关系的字段或字符，例如，姓氏、身份证号码中的预定位数的数值（如身份证号码的前 14 位数字）、社会保障卡号码或其它能够确定用户身份或信息的证件号码等。

在实施中，可以通过多种方式获取用户数据，例如，可以通过购买的方式从不同的用户处购买其用户数据，或者，用户注册某网站或应用程序时填写的信息，如注册支付宝时填写的信息等，或者，用户主动上传的用户数据等，其中，具体通过何种方式获取用户数据，本申请实施例对此不做限定。获取到用户数据后，可以将获取的用户数据中包含的数据字段进行对比，从中查找出其数据字段有相同的部分的用户数据，并可以将数据字段中有相同的部分的用户数据组成一组，得到一个用户数据对，通过上述方式，可以得到多组用户数据对，且每个用户数据对中都包含有数据字段的相同部分。

例如，在实际应用中，为了尽可能的减少运算量、提高处理效率，可以设定数据字段为身份证号码和姓氏，则可以在用户数据中查找用户的身份证号码和姓名等信息，考虑到身份证号码的某一位数字或多位数字可以表征两个用户之间的关系，例如，身份证号码的前 14 位数字等。本申请实施例中以身份证

号码的前 14 位数字作为判定数据字段中是否具有相同部分的依据为例，具体地，可以获取每个用户的身份证号码的前 14 位数字和姓氏，并比较不同用户的身份证号码的前 14 位数字和姓氏。可以将具有相同姓氏且身份证号码的前14 位数字相同的两组用户数据划分到同一个用户数据对中。具体可以通过用户对的形式存储用户数据对，例如，{用户 1 身份证号码，用户 2 身份证号码，用户 1 姓名，用户 2 姓名，用户 1 其它数据，用户 2 其它数据}等。

需要说明的是，上述两组用户数据的数据字段有相同的部分可以理解为数据字段中的一部分内容相同，如上述内容中 18 位身份证号码的前 14 位数字等，也可以理解为数据字段的全部内容相同等。

在步骤 S104 中，获取每个用户数据对所对应的用户相似度，用户相似度为每个用户数据对中的两组用户数据对应的用户之间的相似度。

其中，用户相似度可以用于表征多个用户之间的相似程度，例如 99%或50%等，在实际应用中，用户相似度还可以通过其它方式表示，例如用户相似度还可以以双胞胎和非双胞胎，或者同卵双胞胎和异卵双胞胎来表示等。

在实施中，本实施例的主要目的是训练分类模型，这样就需要训练分类模型的样本数据，以及该样本数据对应的用户相似度，用户相似度可以预先存储于服务器或终端设备中。用户相似度的确定可以包括多种方式，以下提供一种可选的处理方式，具体可以参见以下内容：可以预先获取用户的图像，该图像可以是用户注册应用程序或网站的过程中用户上传的，其中的用户可以是每个用户数据对中包含的两组用户数据对应的用户。可以将每个用户数据对中的图像进行对比，通过图像的对比，可以计算该用户数据对中的两组用户数据对应的用户之间的相似度。在进行图像对比的过程中，可以使用如图像预处理、图像特征提取、图像特征对比等处理方式，本申请实施例对此不做限定。

在步骤 S106 中，根据每个用户数据对所对应的用户相似度和上述多个用户数据对，确定用于训练预设的分类模型的样本数据。

其中，分类模型可以是任意分类模型，如朴素贝叶斯分类模型、Logistic回归分类模型、决策树分类模型或支持向量机分类模型等，本申请实施例中考虑到分类模型仅用于判断两个不同用户之间是否相似，因此，该分类模型可以选用二分类模型。样本数据可以是用于训练分类模型的数据，该样本数据可以是用户数据对中的两组用户数据，也可以是上述用户数据经过某种处理后得到的数据等，如对上述用户数据进行特征提取，得到相应的用户特征，该用户特

征的数据可以作为样本数据。

在实施中，可以预先设置相似度阈值，如80%或70%等，然后，可以将每个用户数据对所对应的用户相似度分别与相似度阈值相比较，可以将用户相似度大于相似度阈值的用户数据对划分为一组，可以将用户相似度小于相似度阈值的用户数据对划分为一组，可以从上述两组中各选取预定数目（如4万或5万等）的用户数据对，并将选取的用户数据对作为用于训练预设的分类模型的样本数据。

需要说明的是，选取用于训练预设的分类模型的样本数据的方式除了上述方式外，还可以包括多种，例如，提取每个用户数据对中包含的两组用户数据的特征，得到相应的用户特征，然后，可以通过每个用户数据对所对应的用户相似度和相似度阈值，将用户特征划分为如上述的两组，可以将两组用户特征的数据作为用于训练预设的分类模型的样本数据。

在步骤S108中，基于上述样本数据对分类模型进行训练，得到相似度分类模型。

其中，相似度分类模型可以是用于确定不同用户之间的相似程度的模型。

在实施中，基于上述选取的用户数据对作为用于训练预设的分类模型的样本数据的情况，可以对选取的用户数据对中的两组用户数据进行特征提取，得到相应的用户特征，然后，可以将样本数据中每个用户数据对的用户特征输入到分类模型中进行计算，计算完成后，可以输出计算结果。可以将该计算结果与相应的用户数据对所对应的用户相似度进行比较，确定两者是否相同，如果两者不同，则可以修改分类模型的相关参数，然后，再将该用户数据对的用户特征输入到修改后的分类模型中进行计算，并判断计算结果与用户相似度是否相同，直到两者相同为止。如果两者相同，则可以选取下一个用户数据对执行上述处理过程，最终每个用户数据对的用户特征输入到分类模型后得到的计算结果与相应的用户数据对所对应的用户相似度均相同，则得到的分类模型即为相似度分类模型。

通过上述方式可以得到相似度分类模型，该相似度分类模型的使用可以参见下述相关内容:

如图2所示，本申请实施例提供一种相似度的确定方法，该方法的执行主体可以为终端设备或服务器，其中的终端设备可以是个人计算机等，服务器可以是独立的一个服务器，也可以是由多个服务器组成的服务器集群。该方法具

体可以包括以下步骤:

在步骤 S202 中，获取待测用户数据对。

其中，待测用户数据对可以是待检测的两个用户的用户数据所组成的用户数据对。

在实施中，为了检测出两个不同用户之间的相似度，可以设置相应的检测应用程序。如图 3 所示，该检测应用程序中可以包括用于上传数据的按键，当需要对两个不同用户进行相似度检测时，可以点击上述用于上传数据的按键，该检测应用程序可以弹出数据上传的提示框，数据上传者可以在提示框中输入待测用户数据对的数据，输入完成后，可以点击该提示框中的确定按键，该检测应用程序可以获取数据上传者输入的待测用户数据对。上述检测应用程序可以安装在终端设备上，也可以安装在服务器上，本申请实施例提供的相似度的确定方法的执行主体若为服务器，且如果检测应用程序安装在终端设备上，则检测应用程序获取到待测用户数据对后，可以将该待测用户数据对发送给服务器，从而服务器可以获取到待测用户数据对。如果检测应用程序安装在服务器上，则服务器通过检测应用程序可以直接获取到待测用户数据对。

在步骤 S204 中，对上述待测用户数据对中每组待测用户数据进行特征提取，得到待测用户特征。

其中，待测用户特征可以是待检测的用户的用户数据的特征。

在实施中，可以获取上述待测用户数据对中每组待测用户数据，针对其中的任意一组待测用户数据，可以使用预先设置的特征提取算法，从该待测用户数据中提取相应的特征，可以将提取的特征作为该待测用户数据对应的待测用户特征。通过上述方式可以得到待测用户数据对中每组待测用户数据对应的待测用户特征。

需要说明的是，特征提取算法可以是能够从用户数据中提取预定特征的任意算法，具体可以根据实际情况进行设定。

在步骤 S206 中，根据上述待测用户特征和预先训练的相似度分类模型，确定上述待测用户数据对中的两组待测用户数据对应的用户之间的相似度。

在实施中，可以将通过上述步骤 S204 得到的待测用户特征输入到通过上述步骤 S102~步骤 S108 得到的相似度分类模型中进行计算，相似度分类模型输出的结果即可以为上述待测用户数据对中的两组待测用户数据对应的用户之间的相似度。

需要说明的是，在实际应用中，相似度分类模型的直接输出结果可以以百分比的方式展现，例如90%或40%等，为了使得输出结果对用户来说更加直观，可以根据实际情况对相似度分类模型的直接输出结果进一步设定，例如，需要区分同卵双胞胎和非同卵双胞胎，或者，需要区分同卵双胞胎和异卵双胞胎等，对于上述情况，可以设置分类阈值，如果直接输出结果大于该分类阈值，则确定上述待测用户数据对中的两组待测用户数据对应的用户之间为同卵双胞胎，否则为非同卵双胞胎或异卵双胞胎等。这样，通过预先训练的相似度分类模型，可以快速判断出待测用户数据对中的两组待测用户数据对应的用户之间的相似度，提高了用户之间相似度的判定效率。

需要说明的是，上述用户数据对和待测用户数据对均是以包含两组用户数据来说明，在实际应用中，本申请提供的模型的训练方法和相似度的确定方法还可以应用于包含两组以上的用户数据的用户数据组合和待测用户数据组合，具体处理可以参见本申请实施例中的相关内容，在此不再赘述。

本申请实施例提供一种模型的训练方法和相似度的确定方法，通过获取的多个用户数据对，且每个用户数据对中的两组用户数据的数据字段有相同的部分，以及获取的每个用户数据对所对应的用户相似度，确定用于训练预设的分类模型的样本数据，然后，基于样本数据对分类模型进行训练，得到相似度分类模型，以便后续可以通过相似度分类模型确定待测用户数据对中的两组待测用户数据对应的用户之间的相似度，这样，仅通过相同的数据字段得到多个用户数据对，并通过用户相似度确定每个用户数据对中的两组用户数据对应的用户之间的关联关系，得到用于训练预设的分类模型的样本数据，而不需要人工标注即可得到样本数据，可以实现模型训练的快速完成，提高了模型训练效率并减少资源消耗。

实施例二

如图4所示，本申请实施例提供了一种数据相似度的确定方法，该方法的执行主体可以为服务器或者该方法可以由终端设备和服务器共同实现，其中的终端设备可以是个人计算机等，服务器可以是独立的一个服务器，也可以是由多个服务器组成的服务器集群。本申请实施例中为了提高模型训练的效率，该方法的执行主体以服务器为例进行详细说明，对于由终端设备和服务器共同实现的按情况，可以参见下述相关内容，在此不再赘述。该方法具体包括如下内

容:

目前人脸识别作为一种用户核实身份的新型方式，在为用户提供便利的同时也产生了新的风险点，目前的人脸识别技术都是利用现场采集的用户图像与该用户在人脸识别系统的数据库中留存的用户图像进行比较，只要比对数值达到预定阈值，则认为该用户为留存的用户图像所对应的用户，以达到核实用户身份的目的。然而，针对长相极为相似的脸，上述方式将很难对用户的身份进行有效核实，从而极易造成因为无法进行身份核实导致的账户误登录以及后续的资金盗用等。

双胞胎特别是同卵双胞胎作为已知的相似脸的最典型情况，因为彼此关系亲密，这样就更容易产生有关负面舆情。如果可以掌握尽可能多的双胞胎用户名单，就可以针对这部分用户群体有单独的人脸识别应对策略以预防上述风险。为此可以构造有效识别双胞胎的模型，在保证高准确率的前提下输出双胞胎名单用于监控这些用户的人脸识别行为以起到风险控制的作用。其中，构造有效识别双胞胎的模型的处理可以参见下述步骤 S402~步骤 S412 提供的模型的训练方法，具体内容如下:

在步骤 S402 中，获取多个用户数据对，其中，每个用户数据对中的两组用户数据的数据字段有相同的部分。

在实施中，考虑到双胞胎通常是姓氏相同且身份证号码的前 14 位数字相同，因此，可以将姓氏和身份证号码的前 14 位数字作为数据字段来选取用户数据对，上述步骤 S402 的具体处理过程可以参见上述实施例一中步骤 S102 的相关内容，在此不再赘述。

需要说明的是，上述选取用户数据对的处理是通过姓氏和身份证号码的前 14 位数字来实现的，在本申请的另一实施例中，选取用户数据对的处理还可以通过其它信息来实现，例如，通过姓氏和社会保障卡号码来实现，或者，通过身份证号码的前 14 位数字和社会保障卡号码来实现等，本申请实施例对此不做限定。

考虑到在对模型进行训练时，需要确定用户数据对中的两组用户数据对应的用户之间的相似程度，以下提供一种相关的处理方式，具体可以参见以下步骤 S404 和步骤 S406。

在步骤 S404 中，获取第一用户数据对所对应的用户的生物特征，其中，第一用户数据对为上述多个用户数据对中的任意用户数据对。

其中，生物特征可以是人体的生理特征和行为特征等，如指纹特征、虹膜特征、面部特征、DNA 等生理特征，再如声纹特征、笔迹特征和击键习惯特征等行为特征。

在实施中，通过上述步骤 S402 的处理获取到多个用户数据对后，可以从多个用户数据对中任意选择一个用户数据对（即第一用户数据对）。用户通过其终端设备登录服务器进行注册时，可以向服务器上传包含该用户上述某一项或多项生物特征，服务器可以将该生物特征与该用户的标识对应存储，其中，用户的标识可以是用户注册时填写的用户名或用户的姓名等，服务器中对应存储的上述信息可以如表 1 所示。

表 1

用户的标识	生物特征
用户 1	生物特征 A
用户 2	生物特征 B
用户 3	生物特征 C

当服务器选取第一用户数据对后，可以从第一用户数据对中分别提取其中包含的用户的标识，然后，通过用户的标识可以获取相应的生物特征，从而得到第一用户数据对所对应的用户的生物特征。例如，第一用户数据对中包含的用户的标识为用户 2 和用户 3，则通过查找如上述表格的对应关系，可以确定用户 2 对应的生物特征为生物特征 B，用户 3 对应的生物特征为生物特征 C，即第一用户数据对所对应的用户的生物特征为生物特征 B 和生物特征 C。

在步骤 S406 中，根据第一用户数据对所对应的用户的生物特征，确定第一用户数据对所对应的用户相似度。

在实施中，通过上述步骤 S404 得到第一用户数据对所对应的用户的生物特征后，可以分别对得到的生物特征进行相似度计算，从而确定相应的两个用户之间的相似程度（即用户相似度），其中，相似度计算可以包括多种实现方式，例如通过特征向量之间的欧氏距离来实现等，本申请实施例对此不做限定。

需要说明的是，可以通过设置阈值来进行相似与否的判断，例如设置阈值为 70，当两个生物特征对应的用户相似度大于 70 时，确定第一用户数据对中的两组用户数据对应的用户相似；当两个生物特征对应的用户相似度小于 70 时，确定第一用户数据对中的两组用户数据对应的用户不相似。

通过上述方式可以对多个用户数据对中除第一用户数据对外的其它用户

数据对执行上述处理过程，从而得到多个用户数据对中每个用户数据对所对应的用户相似度。

上述步骤 S404 和步骤 S406 是通过用户的生物特征确定用户相似度的，在实际应用中，确定用户相似度具体可以通过多种实现方式实现，以下以生物特征为面部特征为例对上述步骤 S404 和步骤 S406 进行具体说明，具体可以参见以下步骤一和步骤二。

步骤一，获取第一用户数据对所对应的用户的面部图像，其中，第一用户数据对为上述多个用户数据对中的任意用户数据对。

在实施中，通过上述步骤 S402 的处理获取到多个用户数据对后，可以从多个用户数据对中任意选择一个用户数据对（即第一用户数据对）。用户通过其终端设备登录服务器进行注册时，可以向服务器上传包含该用户面部的图像，服务器可以将该图像与该用户的标识对应存储，其中，用户的标识可以是用户注册时填写的用户名或用户的姓名等，服务器中对应存储的上述信息可以如表 2 所示。

表 2

用户的标识	包含用户面部的图像
用户 1	图像 A
用户 2	图像 B
用户 3	图像 C

当服务器选取第一用户数据对后，可以从第一用户数据对中分别提取其中包含的用户的标识，然后，通过用户的标识可以获取相应的图像，从而得到第一用户数据对所对应的用户的面部图像。例如，第一用户数据对中包含的用户的标识为用户 2 和用户 3，则通过查找如上述表格的对应关系，可以确定用户 2 对应的包含用户面部的图像为图像 B，用户 3 对应的包含用户面部的图像为图像 C，即第一用户数据对所对应的用户的面部图像为图像 B 和图像 C。

步骤二，对上述面部图像进行特征提取，得到面部图像特征，并根据第一用户数据对所对应的用户的面部特征，确定第一用户数据对所对应的用户相似度。

在实施中，通过上述步骤一得到第一用户数据对所对应的用户的面部图像后，可以分别对得到的面部图像进行特征提取，得到相应的面部图像特征，并基于每个面部图像的提取特征得到相应的特征向量，然后，可以计算其中任意

两个面部图像的特征向量之间的欧式距离，通过特征向量之间的欧式距离的数值大小，可以确定相应的两个用户之间的相似程度（即用户相似度），其中，特征向量之间的欧式距离的数值越大，用户相似度越低；特征向量之间的欧式距离的数值越小，用户相似度越高。

需要说明的是，对于面部图像而言，两个面部图像只有相似和非相似的区别，为此，可以通过设置阈值来进行相似与否的判断，例如设置阈值为 70，当两个面部图像对应的用户相似度大于 70 时，确定第一用户数据对中的两组用户数据对应的用户相似；当两个面部图像对应的用户相似度小于 70 时，确定第一用户数据对中的两组用户数据对应的用户不相似。

例如，基于上述步骤一的示例，分别对图像 B 和图像 C 进行特征提取，通过提取的特征分别构建相应的特征向量，得到图像 B 的特征向量和图像 C 的特征向量。计算图像 B 的特征向量和图像 C 的特征向量之间的欧式距离，通过得到的欧式距离的数值确定用户 2 和用户 3 之间的用户相似度。

通过上述方式可以对多个用户数据对中除第一用户数据对外的其它用户数据对执行上述处理过程，从而得到多个用户数据对中每个用户数据对所对应的用户相似度。

此外，对于上述步骤 S404 和步骤 S406 的处理，以下再提供一种可选的处理方式，具体可以参见以下步骤一和步骤二。

步骤一，获取第一用户数据对所对应的用户的语音数据，其中，第一用户数据对为多个用户数据对中的任意用户数据对。

在实施中，通过上述步骤 S402 的处理获取到多个用户数据对后，可以从多个用户数据对中任意选择一个用户数据对（即第一用户数据对）。用户通过其终端设备登录服务器进行注册时，可以向服务器上传包含预定时长（如 3 秒或 5 秒等）和/或预定语音内容（如一个或多个词的语音或一句话的语音等）的语音数据，服务器可以将该语音数据与该用户的标识对应存储。当服务器选取第一用户数据对后，可以从第一用户数据对中分别提取其中包含的用户的标识，然后，通过用户的标识可以获取相应的语音数据，从而得到第二用户数据对所对应的用户的语音数据。

步骤二，对上述语音数据进行特征提取，得到语音特征，并根据第一用户数据对所对应的用户的语音特征，确定第一用户数据对所对应的用户相似度。

在实施中，通过上述步骤一得到第一用户数据对所对应的用户的语音数据

后，可以分别对得到的语音数据进行特征提取，并基于每个语音数据的提取特征确定相应的两个用户之间的相似程度（即用户相似度），具体处理过程可以参见上述步骤 S406 中的相关内容，或者，可以通过特征的逐一比对的方式确定用户相似度，又或者，可以对任意两个语音数据进行语音频谱分析，以确定用户相似度等。通过上述方式可以对多个用户数据对中除第一用户数据对外的其它用户数据对执行上述处理过程，从而得到多个用户数据对中每个用户数据对所对应的用户相似度。

在步骤 S408 中，对上述多个用户数据对中的每个用户数据对进行特征提取，得到每个用户数据对中的两组用户数据之间相关联的用户特征。

在实施中，可以从多个用户数据对中任意选取一个用户数据对（可以称为第三用户数据对），可以对第三用户数据对中的两组不同的用户数据分别进行特征提取，例如，第三用户数据对中包括用户数据 1 和用户数据 2，可以对用户数据 1 进行特征提取，并对用户数据 2 进行特征提取。然后，可以对比在不同的用户数据中提取的特征，从而得到第三用户数据对中的两组用户数据之间相关联的用户特征。通过上述方式可以对多个用户数据对中除第三用户数据对外的其它用户数据对执行上述处理过程，从而得到每个用户数据对中的两组用户数据之间相关联的用户特征。

在实际应用中，用户特征可以包含但不限于户籍维度特征、姓名维度特征、社交特征和兴趣爱好特征等特征。其中，户籍维度特征可以包括用户身份信息的特征。户籍维度特征主要是基于中国的户籍管理制度，户籍中包含的身份证信息中包括出生日期和户籍申报地，同时户籍中具有父母姓名和公民住址，然而由于历史和其它原因，部分公民登记的信息并不与实际情况一样，存在如提前申报生日、双方分别随父母姓，甚至父母离异导致户籍分离等情况，所以户籍维度特征对于判定两个用户是否为双胞胎起到一定的参考作用。这样，通过用户数据对所对应的不同用户之间的出生日期是否一致、户籍申报地是否一致、是否有共同父母、现住址的一致程度等特征确定不同用户之间的关联。

姓名维度特征包括用户姓名信息的特征和用户姓氏的稀缺程度的特征。对于姓名维度特征，基于 NLP（Nature Language Processing，自然语言处理）理论和社会经验，通常，如果两个人的名字看起来比较像，比如张金龙和张金虎，或者具有某种语义关联，如张美美和张丽丽，则认为两者之间应该具有某种关联。在本申请实施例中，可以引入词典来评估两个用户在名字上的关系，同时

利用用户注册的个人信息和人口统计数据统计姓氏的稀缺程度作为特征。这样，通过用户数据对所对应的不同用户之间的姓氏是否一致、姓名长度是否一致、名字近义词程度、名字组合是否为词和姓氏稀缺程度等特征确定不同用户之间的关联。

社交特征包括用户的社会关系信息的特征。对于社交特征，可以是基于大数据对用户数据对的社会关系进行提炼而成，通常，双胞胎应该具有较多的互动和重复性较高的社会关系，如共同的亲戚，甚至同学等。在本申请实施例中，基于服务器中存储的用户的个人信息构成的关系网络、通讯录等已有数据对用户数据对进行关联，以得到相应的特征。这样，通过用户数据对所对应的不同用户之间的社交应用是否互相关注、是否有资金往来、通讯录中是否包含对方的联系方式、通讯录标注是否有称谓和通讯录的交集数量等特征确定不同用户之间的关联。

此外，考虑到双胞胎具有较多的共同爱好、购物兴趣，以及可能会共同出游等，用户特征还可以包括如电商、旅游、文娱等多维度特征，在本申请实施例中，电商、旅游、文娱等多维度特征的相关数据可以从预定的数据库或某网站中获取得到。这样，通过用户数据对所对应的不同用户之间的购物记录的交集数量、是否有过同时出游、是否同时入住过酒店、购物倾向的相似度和收货地址是否一样等特征确定不同用户之间的关联。

需要说明的是，上述确定用户相似度的处理（即包括步骤 S404 和步骤 S406）和特征提取的处理（即步骤 S408）是按照先后顺序执行的，在实际应用中，确定用户相似度的处理和特征提取的处理可以同时执行，也可以先执行特征提取的处理，然后再执行确定用户相似度的处理，本申请实施例对此不做限定。

在步骤 S410 中，根据每个用户数据对中的两组用户数据之间相关联的用户特征和每个用户数据对所对应的用户相似度，确定用于训练分类模型的样本数据。

在实施中，可以预先设置阈值，通过阈值可以从多个用户数据对中选取用户相似度大于该阈值的用户数据对，可以将选取的用户数据对中的两组用户数据之间相关联的用户特征作为训练分类模型的用户特征，可以将选取的用户特征和选取的用户数据对所对应的用户相似度确定为用于训练分类模型的样本数据。

上述步骤 S410 的处理除了可以采用上述方式外，还可以采用多种方式处

理，以下还提供一种可选的处理方式，具体可以包括以下步骤一和步骤二:

步骤一，根据每个用户数据对所对应的用户相似度和预定的相似度阈值，从多个用户数据对所对应的用户特征中选取正样本特征和负样本特征。

在实施中，基于同卵双胞胎长相高度相似这一常识，以及双胞胎出生日期、出生地等相同，且通常情况下双胞胎的姓氏也相同的社会常识，通过两个用户的面部图像计算用户相似度，从而确定两个用户是否为同卵双胞胎，具体地，可以预先设置相似度阈值，如80%或70%等，可以将用户相似度大于相似度阈值的用户数据对确定为同卵双胞胎的用户数据对，可以将用户相似度小于相似度阈值的用户数据对确定为非同卵双胞胎的用户数据对。同时，由于同卵双胞胎和异卵双胞胎除了在长相上有所差异外，其它特征基本一致，所以，可以将同卵双胞胎的用户数据对所对应的用户特征作为相似度分类模型的正样本特征，而非同卵双胞胎（包括异卵双胞胎和非双胞胎）的用户数据对所对应的用户特征则作为相似度分类模型的负样本特征。

需要说明的是，负样本特征并不是指其中包含的特征全部都是异卵双胞胎的用户特征，在实际应用中，异卵双胞胎的用户特征也可能在负样本特征中的比例极少，还可能在负样本特征中包含有少量的正样本特征，而这样并不会影响分类模型的训练，反而会有助于提升相似度分类模型的鲁棒性。

此外，正样本特征和负样本特征中包含的特征数目可以相同。例如，从多个用户数据对中选取用户相似度小于10%的10000个用户数据对，从多个用户数据对中选取用户相似度大于10%且小于20%的10000个用户数据对，从多个用户数据对中选取用户相似度大于20%且小于30%的10000个用户数据对，从多个用户数据对中选取用户相似度大于30%且小于40%的10000个用户数据对，从多个用户数据对中选取用户相似度大于40%且小于50%的10000个用户数据对，将上述50000个用户数据对的用户特征作为负样本特征。从多个用户数据对中选取用户相似度大于80%且小于90%的40000个用户数据对，从多个用户数据对中选取用户相似度大于90%且小于100%的10000个用户数据对，将上述50000个用户数据对的用户特征作为正样本特征。

步骤二，将正样本特征和负样本特征作为用于训练分类模型的样本数据。

在实施中，可以将用户特征和相应的用户相似度的数据组合，可以将组合后的数据作为用于训练分类模型的样本数据。

在步骤S412中，基于样本数据对分类模型进行训练，得到相似度分类模

型。

其中，由于分类模型的主要目的是识别出双胞胎，因此，为了使得本申请实施例简化可行，相似度分类模型可以为二分类器模型，具体如 GBDT（Gradient Boosting Decision Tree，迭代决策树）二分类器模型。

在实施中，可以分别将正样本特征输入到分类模型中进行计算，得到的计算结果可以与该正样本特征相应的用户相似度对比，如果两者相匹配，则可以选择下一个正样本特征或负样本特征输入到分类模型中进行计算。得到的计算结果继续与该正样本特征相应的用户相似度匹配对比。如果两者不匹配，则可以调整分类模型中的相关参数的数值，然后再将该正样本特征输入到分类模型中进行计算，得到的计算结果再与该正样本特征相应的用户相似度匹配对比，即重复上述过程，直到两者相匹配为止。通过上述方式，可以将所有的正样本特征和负样本特征输入到分类模型中进行计算，从而达到对分类模型进行训练的目的，可以将最终训练得到的分类模型作为相似度分类模型。

通过上述处理过程得到了相似度分类模型，该相似度分类模型可以用于人脸识别场景中，对于具有风险的双胞胎用户，通过该相似度分类模型可以进行单独的风险控制。

得到相似度分类模型后，可以应用相似度分类模型来判定待测用户数据对所对应的待测用户是否为双胞胎，如图 5 所示，其中的具体处理可以参见以下步骤 S414~步骤 S420 的内容。

在步骤 S414 中，获取待测用户数据对。

上述步骤 S414 的步骤内容与上述实施例一中步骤 S202 的步骤内容相同，步骤 S414 的具体处理可以参见步骤 S202 的相关内容，在此不再赘述。

在步骤 S416 中，对待测用户数据对中每组待测用户数据进行特征提取，得到待测用户特征。

其中上述步骤 S416 中对待测用户数据对中每组待测用户数据进行特征提取，得到待测用户特征的处理过程，可以参见上述步骤 S408 的相关内容，即从待测用户数据中提取的特征包括但不限于籍维度特征、姓名维度特征、社交特征和兴趣爱好特征等，参见上述步骤 S408 的相关内容，在此不再赘述。

在步骤 S418 中，根据待测用户特征和预先训练的相似度分类模型，确定待测用户数据对中的两组待测用户数据对应的用户之间的相似度。

上述步骤 S418 的步骤内容与上述实施例一中步骤 S206 的步骤内容相同，

步骤 S418 的具体处理可以参见步骤 S206 的相关内容，在此不再赘述。

在步骤 S420 中，如果待测用户数据对中的两组待测用户数据对应的用户之间的相似度大于预定相似度分类阈值，则确定待测用户数据对所对应的待测用户为双胞胎。

在实施中，由于输出的双胞胎名单会影响目标用户的人脸识别功能的使用，因此，使用的过程中需要追求相似度分类模型的高准确度，在实际应用中可以设置一个较大的数值作为相似度分类阈值，例如，95%作为相似度分类阈值或97%作为相似度分类阈值等。利用训练好的相似度分类模型对待测用户特征进行预测并输出评分。其中，评分过程是计算相应的用户数据对所对应的用户为双胞胎的概率，比如概率为 80%，则评分为 80 分，概率为 90%，则评分为 90 分，得到的分数越高，用户数据对所对应的用户为双胞胎的概率越高。

本申请实施例提供一种数据相似度的确定方法，通过获取的多个用户数据对，且每个用户数据对中的两组用户数据的数据字段有相同的部分，以及获取的每个用户数据对所对应的用户相似度，确定用于训练预设的分类模型的样本数据，然后，基于样本数据对分类模型进行训练，得到相似度分类模型，以便后续可以通过相似度分类模型确定待测用户数据对中的两组待测用户数据对应的用户之间的相似度，这样，仅通过相同的数据字段得到多个用户数据对，并通过用户相似度确定每个用户数据对中的两组用户数据对应的用户之间的关联关系，得到用于训练预设的分类模型的样本数据，而不需要人工标注即可得到样本数据，可以实现模型训练的快速完成，提高了模型训练效率并减少资源消耗。

实施例三

以上为本申请实施例提供的数据相似度的确定方法，基于同样的思路，本申请实施例还提供一种模型的训练装置，如图 6 所示。

所述模型的训练装置可以设置在服务器中，该装置包括：数据获取模块 601、相似度获取模块 602、样本数据确定模块 603 和模型训练模块 604，其中：

数据获取模块 601，用于获取多个用户数据对，其中，所述每个用户数据对中的两组用户数据的数据字段有相同的部分；

相似度获取模块 602，用于获取每个用户数据对所对应的用户相似度，所述用户相似度为每个用户数据对中的两组用户数据对应的用户之间的相似度；

样本数据确定模块 603，用于根据所述每个用户数据对所对应的用户相似度和所述多个用户数据对，确定用于训练预设的分类模型的样本数据；

模型训练模块 604，用于基于所述样本数据对所述分类模型进行训练，得到相似度分类模型。

本申请实施例中，所述相似度获取模块 602，包括：

生物特征获取单元，用于获取第一用户数据对所对应的用户的生物特征，其中，所述第一用户数据对为所述多个用户数据对中的任意用户数据对；

相似度获取单元，用于根据所述第一用户数据对所对应的用户的生物特征，确定所述第一用户数据对所对应的用户相似度。

本申请实施例中，所述生物特征包括面部图像特征，

所述生物特征获取单元，用于获取第一用户数据对所对应的用户的面部图像；对所述面部图像进行特征提取，得到面部图像特征；

相应的，所述相似度获取单元，用于根据所述第一用户数据对所对应的用户的面部图像特征，确定所述第一用户数据对所对应的用户相似度。

本申请实施例中，所述生物特征包括语音特征，

所述生物特征获取单元，用于获取第一用户数据对所对应的用户的语音数据；对所述语音数据进行特征提取，得到语音特征；

相应的，所述相似度获取单元，用于根据所述第一用户数据对所对应的用户的语音特征，确定所述第一用户数据对所对应的用户相似度。

本申请实施例中，所述样本数据确定模块 603，包括：

特征提取单元，用于对所述多个用户数据对中的每组用户数据对进行特征提取，得到每个用户数据对中的两组用户数据之间相关联的用户特征；

样本数据确定单元，用于根据所述每个用户数据对中的两组用户数据之间相关联的用户特征和所述每个用户数据对所对应的用户相似度，确定用于训练分类模型的样本数据。

本申请实施例中，所述样本数据确定单元，用于根据每个用户数据对所对应的用户相似度和预定的相似度阈值，从所述多个用户数据对所对应的用户特征中选取正样本特征和负样本特征；将所述正样本特征和负样本特征作为用于训练分类模型的样本数据。

本申请实施例中，所述用户特征包括户籍维度特征、姓名维度特征、社交特征和兴趣爱好特征；所述户籍维度特征包括用户身份信息的特征，所述姓名

维度特征包括用户姓名信息的特征和用户姓氏的稀缺程度的特征，所述社交特征包括用户的社会关系信息的特征。

本申请实施例中，所述正样本特征和负样本特征中包含的特征数目相同。

本申请实施例中，所述相似度分类模型为二分类器模型。

本申请实施例提供一种模型的训练装置，通过获取的多个用户数据对，且每个用户数据对中的两组用户数据的数据字段有相同的部分，以及获取的每个用户数据对所对应的用户相似度，确定用于训练预设的分类模型的样本数据，然后，基于样本数据对分类模型进行训练，得到相似度分类模型，以便后续可以通过相似度分类模型确定待测用户数据对中的两组待测用户数据对应的用户之间的相似度，这样，仅通过相同的数据字段得到多个用户数据对，并通过用户相似度确定每个用户数据对中的两组用户数据对应的用户之间的关联关系，得到用于训练预设的分类模型的样本数据，而不需要人工标注即可得到样本数据，可以实现模型训练的快速完成，提高了模型训练效率并减少资源消耗。

实施例四

以上为本申请实施例提供的模型的训练装置，基于同样的思路，本申请实施例还提供一种数据相似度的确定装置，如图 7 所示。

所述数据相似度的确定装置包括：待测数据获取模块 701、特征提取模块 702 和相似度确定模块 703，其中：

待测数据获取模块 701，用于获取待测用户数据对；

特征提取模块 702，用于对所述待测用户数据对中每组待测用户数据进行特征提取，得到待测用户特征；

相似度确定模块 703，用于根据所述待测用户特征和预先训练的相似度分类模型，确定所述待测用户数据对中的两组待测用户数据对应的用户之间的相似度。

本申请实施例中，所述装置还包括：

相似度分类模块，用于如果所述待测用户数据对中的两组待测用户数据对应的用户之间的相似度大于预定相似度分类阈值，则确定所述待测用户数据对所对应的待测用户为双胞胎。

本申请实施例提供一种数据相似度的确定装置，通过获取的多个用户数据对，且每个用户数据对中的两组用户数据的数据字段有相同的部分，以及获取

的每个用户数据对所对应的用户相似度，确定用于训练预设的分类模型的样本数据，然后，基于样本数据对分类模型进行训练，得到相似度分类模型，以便后续可以通过相似度分类模型确定待测用户数据对中的两组待测用户数据对应的用户之间的相似度，这样，仅通过相同的数据字段得到多个用户数据对，并通过用户相似度确定每个用户数据对中的两组用户数据对应的用户之间的关联关系，得到用于训练预设的分类模型的样本数据，而不需要人工标注即可得到样本数据，可以实现模型训练的快速完成，提高了模型训练效率并减少资源消耗。

实施例五

基于同样的思路，本申请实施例还提供一种模型的训练设备，如图 8 所示。

该模型的训练设备可以为上述实施例提供的服务器等。

模型的训练设备可因配置或性能不同而产生比较大的差异，可以包括一个或一个以上的处理器 801 和存储器 802，存储器 802 中可以存储有一个或一个以上存储应用程序或数据。其中，存储器 802 可以是短暂存储或持久存储。存储在存储器 802 的应用程序可以包括一个或一个以上模块（图示未示出），每个模块可以包括对模型的训练设备中的一系列计算机可执行指令。更进一步地，处理器 801 可以设置为与存储器 802 通信，在模型的训练设备上执行存储器 802 中的一系列计算机可执行指令。模型的训练设备还可以包括一个或一个以上电源 803，一个或一个以上有线或无线网络接口 804，一个或一个以上输入输出接口 805，一个或一个以上键盘 806。

具体在本实施例中，模型的训练设备包括有存储器，以及一个或一个以上的程序，其中一个或者一个以上程序存储于存储器中，且一个或者一个以上程序可以包括一个或一个以上模块，且每个模块可以包括对模型的训练设备中的一系列计算机可执行指令，且经配置以由一个或者一个以上处理器执行该一个或者一个以上程序包含用于进行以下计算机可执行指令：

获取多个用户数据对，其中，所述每个用户数据对中的两组用户数据的数据字段有相同的部分；

获取每个用户数据对所对应的用户相似度，所述用户相似度为每个用户数据对中的两组用户数据对应的用户之间的相似度；

根据所述每个用户数据对所对应的用户相似度和所述多个用户数据对，确

定用于训练预设的分类模型的样本数据;

基于所述样本数据对所述分类模型进行训练，得到相似度分类模型。

可选地，所述可执行指令在被执行时，还可以使所述处理器:

获取第一用户数据对所对应的用户的生物特征，其中，所述第一用户数据对为所述多个用户数据对中的任意用户数据对;

根据所述第一用户数据对所对应的用户的生物特征，确定所述第一用户数据对所对应的用户相似度。

可选地，所述可执行指令在被执行时，还可以使所述处理器:

所述生物特征包括面部图像特征，

所述获取第一用户数据对所对应的用户的生物特征，包括:

获取第一用户数据对所对应的用户的面部图像;

对所述面部图像进行特征提取，得到面部图像特征;

相应的，所述根据所述第一用户数据对所对应的用户的生物特征，确定所述第一用户数据对所对应的用户相似度，包括:

根据所述第一用户数据对所对应的用户的面部图像特征，确定所述第一用户数据对所对应的用户相似度。

可选地，所述可执行指令在被执行时，还可以使所述处理器:

所述生物特征包括语音特征，

所述获取第一用户数据对所对应的用户的生物特征，包括:

获取第一用户数据对所对应的用户的语音数据;

对所述语音数据进行特征提取，得到语音特征;

相应的，所述根据所述第一用户数据对所对应的用户的生物特征，确定所述第一用户数据对所对应的用户相似度，包括:

根据所述第一用户数据对所对应的用户的语音特征，确定所述第一用户数据对所对应的用户相似度。

可选地，所述可执行指令在被执行时，还可以使所述处理器:

对所述多个用户数据对中的每个用户数据对进行特征提取，得到每个用户数据对中的两组用户数据之间相关联的用户特征;

根据所述每个用户数据对中的两组用户数据之间相关联的用户特征和所述每个用户数据对所对应的用户相似度，确定用于训练分类模型的样本数据。

可选地，所述可执行指令在被执行时，还可以使所述处理器:

根据每个用户数据对所对应的用户相似度和预定的相似度阈值，从所述多个用户数据对所对应的用户特征中选取正样本特征和负样本特征；

将所述正样本特征和负样本特征作为用于训练分类模型的样本数据。

可选地，所述用户特征包括户籍维度特征、姓名维度特征、社交特征和兴趣爱好特征；所述户籍维度特征包括用户身份信息的特征，所述姓名维度特征包括用户姓名信息的特征和用户姓氏的稀缺程度的特征，所述社交特征包括用户的社会关系信息的特征。

可选地，所述正样本特征和负样本特征中包含的特征数目相同。

可选地，所述相似度分类模型为二分类器模型。

本申请实施例提供一种模型的训练设备，通过获取的多个用户数据对，且每个用户数据对中的两组用户数据的数据字段有相同的部分，以及获取的每个用户数据对所对应的用户相似度，确定用于训练预设的分类模型的样本数据，然后，基于样本数据对分类模型进行训练，得到相似度分类模型，以便后续可以通过相似度分类模型确定待测用户数据对中的两组待测用户数据对应的用户之间的相似度，这样，仅通过相同的数据字段得到多个用户数据对，并通过用户相似度确定每个用户数据对中的两组用户数据对应的用户之间的关联关系，得到用于训练预设的分类模型的样本数据，而不需要人工标注即可得到样本数据，可以实现模型训练的快速完成，提高了模型训练效率并减少资源消耗。

实施例六

基于同样的思路，本申请实施例还提供一种数据相似度的确定设备，如图9所示。

该数据相似度的确定设备可以为上述实施例提供的服务器或终端设备等。

数据相似度的确定设备可因配置或性能不同而产生比较大的差异，可以包括一个或一个以上的处理器901和存储器902，存储器902中可以存储有一个或一个以上存储应用程序或数据。其中，存储器902可以是短暂存储或持久存储。存储在存储器902的应用程序可以包括一个或一个以上模块(图示未示出)，每个模块可以包括对数据相似度的确定设备中的一系列计算机可执行指令。更进一步地，处理器901可以设置为与存储器902通信，在数据相似度的确定设备上执行存储器902中的一系列计算机可执行指令。数据相似度的确定设备还可以包括一个或一个以上电源903，一个或一个以上有线或无线网络接口904，

一个或一个以上输入输出接口 905，一个或一个以上键盘 906。

具体在本实施例中，数据相似度的确定设备包括有存储器，以及一个或一个以上的程序，其中一个或者一个以上程序存储于存储器中，且一个或者一个以上程序可以包括一个或一个以上模块，且每个模块可以包括对数据相似度的确定设备中的一系列计算机可执行指令，且经配置以由一个或者一个以上处理器执行该一个或者一个以上程序包含用于进行以下计算机可执行指令：

获取待测用户数据对；

对所述待测用户数据对中每组待测用户数据进行特征提取，得到待测用户特征；

根据所述待测用户特征和预先训练的相似度分类模型，确定所述待测用户数据对中的两组待测用户数据对应的用户之间的相似度。

可选地，所述可执行指令在被执行时，还可以使所述处理器：

如果所述待测用户数据对中的两组待测用户数据对应的用户之间的相似度大于预定相似度分类阈值，则确定所述待测用户数据对所对应的待测用户为双胞胎。

本申请实施例提供一种数据相似度的确定设备，通过获取的多个用户数据对，且每个用户数据对中的两组用户数据的数据字段有相同的部分，以及获取的每个用户数据对所对应的用户相似度，确定用于训练预设的分类模型的样本数据，然后，基于样本数据对分类模型进行训练，得到相似度分类模型，以便后续可以通过相似度分类模型确定待测用户数据对中的两组待测用户数据对应的用户之间的相似度，这样，仅通过相同的数据字段得到多个用户数据对，并通过用户相似度确定每个用户数据对中的两组用户数据对应的用户之间的关联关系，得到用于训练预设的分类模型的样本数据，而不需要人工标注即可得到样本数据，可以实现模型训练的快速完成，提高了模型训练效率并减少资源消耗。

上述对本说明书特定实施例进行了描述。其它实施例在所附权利要求书的范围内。在一些情况下，在权利要求书中记载的动作或步骤可以按照不同于实施例中的顺序来执行并且仍然可以实现期望的结果。另外，在附图中描绘的过程不一定要求示出的特定顺序或者连续顺序才能实现期望的结果。在某些实施方式中，多任务处理和并行处理也是可以的或者可能是有利的。

在 20 世纪 90 年代，对于一个技术的改进可以很明显地区分是硬件上的改

进（例如，对二极管、晶体管、开关等电路结构的改进）还是软件上的改进（对于方法流程的改进）。然而，随着技术的发展，当今的很多方法流程的改进已经可以视为硬件电路结构的直接改进。设计人员几乎都通过将改进的方法流程编程到硬件电路中来得到相应的硬件电路结构。因此，不能说一个方法流程的改进就不能用硬件实体模块来实现。例如，可编程逻辑器件（Programmable Logic Device，PLD）（例如现场可编程门阵列（Field Programmable Gate Array，FPGA））就是这样一种集成电路，其逻辑功能由用户对器件编程来确定。由设计人员自行编程来把一个数字系统“集成”在一片 PLD 上，而不需要请芯片制造厂商来设计和制作专用的集成电路芯片。而且，如今，取代手工地制作集成电路芯片，这种编程也多半改用“逻辑编译器（logic compiler）”软件来实现，它与程序开发撰写时所用的软件编译器相类似，而要编译之前的原始代码也得用特定的编程语言来撰写，此称之为硬件描述语言（Hardware Description Language，HDL），而 HDL 也并非仅有一种，而是有许多种，如 ABEL（Advanced Boolean Expression Language）、AHDL（Altera Hardware Description Language）、Confluence、CUPL（Cornell University Programming Language）、HDCal、JHDL（Java Hardware Description Language）、Lava、Lola、MyHDL、PALASM、RHDL（Ruby Hardware Description Language）等，目前最普遍使用的是 VHDL（Very-High-Speed Integrated Circuit Hardware Description Language）与 Verilog。本领域技术人员也应该清楚，只需要将方法流程用上述几种硬件描述语言稍作逻辑编程并编程到集成电路中，就可以很容易得到实现该逻辑方法流程的硬件电路。

控制器可以按任何适当的方式实现，例如，控制器可以采取例如微处理器或处理器以及存储可由该（微）处理器执行的计算机可读程序代码（例如软件或固件）的计算机可读介质、逻辑门、开关、专用集成电路（Application Specific Integrated Circuit，ASIC）、可编程逻辑控制器和嵌入微控制器的形式，控制器的例子包括但不限于以下微控制器：ARC 625D、Atmel AT91SAM、Microchip PIC18F26K20 以及 Silicone Labs C8051F320，存储器控制器还可以被实现为存储器的控制逻辑的一部分。本领域技术人员也知道，除了以纯计算机可读程序代码方式实现控制器以外，完全可以通过将方法步骤进行逻辑编程来使得控制器以逻辑门、开关、专用集成电路、可编程逻辑控制器和嵌入微控制器等的形式来实现相同功能。因此这种控制器可以被认为是一种硬件部件，而对其内包

括的用于实现各种功能的装置也可以视为硬件部件内的结构。或者甚至，可以将用于实现各种功能的装置视为既可以是实现方法的软件模块又可以是硬件部件内的结构。

上述实施例阐明的系统、装置、模块或单元，具体可以由计算机芯片或实体实现，或者由具有某种功能的产品来实现。一种典型的实现设备为计算机。具体的，计算机例如可以为个人计算机、膝上型计算机、蜂窝电话、相机电话、智能电话、个人数字助理、媒体播放器、导航设备、电子邮件设备、游戏控制台、平板计算机、可穿戴设备或者这些设备中的任何设备的组合。

为了描述的方便，描述以上装置时以功能分为各种单元分别描述。当然，在实施本申请时可以把各单元的功能在同一个或多个软件和/或硬件中实现。

本领域内的技术人员应明白，本申请的实施例可提供为方法、系统、或计算机程序产品。因此，本申请可采用完全硬件实施例、完全软件实施例、或结合软件和硬件方面的实施例的形式。而且，本申请可采用在一个或多个其中包含有计算机可用程序代码的计算机可用存储介质（包括但不限于磁盘存储器、CD-ROM、光学存储器等）上实施的计算机程序产品的形式。

本申请是参照根据本申请实施例的方法、设备（系统）、和计算机程序产品的流程图和／或方框图来描述的。应理解可由计算机程序指令实现流程图和／或方框图中的每一流程和／或方框、以及流程图和／或方框图中的流程和／或方框的结合。可提供这些计算机程序指令到通用计算机、专用计算机、嵌入式处理机或其他可编程数据处理设备的处理器以产生一个机器，使得通过计算机或其他可编程数据处理设备的处理器执行的指令产生用于实现在流程图一个流程或多个流程和／或方框图一个方框或多个方框中指定的功能的装置。

这些计算机程序指令也可存储在能引导计算机或其他可编程数据处理设备以特定方式工作的计算机可读存储器中，使得存储在该计算机可读存储器中的指令产生包括指令装置的制造品，该指令装置实现在流程图一个流程或多个流程和／或方框图一个方框或多个方框中指定的功能。

这些计算机程序指令也可装载到计算机或其他可编程数据处理设备上，使得在计算机或其他可编程设备上执行一系列操作步骤以产生计算机实现的处理，从而在计算机或其他可编程设备上执行的指令提供用于实现在流程图一个流程或多个流程和／或方框图一个方框或多个方框中指定的功能的步骤。

在一个典型的配置中，计算设备包括一个或多个处理器（CPU）、输入/输

出接口、网络接口和内存。

内存可能包括计算机可读介质中的非永久性存储器，随机存取存储器（RAM）和/或非易失性内存等形式，如只读存储器（ROM）或闪存（flash RAM）。内存是计算机可读介质的示例。

计算机可读介质包括永久性和非永久性、可移动和非可移动媒体可以由任何方法或技术来实现信息存储。信息可以是计算机可读指令、数据结构、程序的模块或其他数据。计算机的存储介质的例子包括，但不限于相变内存（PRAM）、静态随机存取存储器（SRAM）、动态随机存取存储器（DRAM）、其他类型的随机存取存储器（RAM）、只读存储器（ROM）、电可擦除可编程只读存储器（EEPROM）、快闪记忆体或其他内存技术、只读光盘只读存储器（CD-ROM）、数字多功能光盘（DVD）或其他光学存储、磁盒式磁带，磁带磁磁盘存储或其他磁性存储设备或任何其他非传输介质，可用于存储可以被计算设备访问的信息。按照本文中的界定，计算机可读介质不包括暂存电脑可读媒体（transitory media），如调制的数据信号和载波。

还需要说明的是，术语“包括”、“包含”或者其任何其他变体意在涵盖非排他性的包含，从而使得包括一系列要素的过程、方法、商品或者设备不仅包括那些要素，而且还包括没有明确列出的其他要素，或者是还包括为这种过程、方法、商品或者设备所固有的要素。在没有更多限制的情况下，由语句“包括一个……”限定的要素，并不排除在包括所述要素的过程、方法、商品或者设备中还存在另外的相同要素。

本领域技术人员应明白，本申请的实施例可提供为方法、系统或计算机程序产品。因此，本申请可采用完全硬件实施例、完全软件实施例或结合软件和硬件方面的实施例的形式。而且，本申请可采用在一个或多个其中包含有计算机可用程序代码的计算机可用存储介质（包括但不限于磁盘存储器、CD-ROM、光学存储器等）上实施的计算机程序产品的形式。

本申请可以在由计算机执行的计算机可执行指令的一般上下文中描述，例如程序模块。一般地，程序模块包括执行特定任务或实现特定抽象数据类型的例程、程序、对象、组件、数据结构等等。也可以在分布式计算环境中实践本申请，在这些分布式计算环境中，由通过通信网络而被连接的远程处理设备来执行任务。在分布式计算环境中，程序模块可以位于包括存储设备在内的本地和远程计算机存储介质中。

本说明书中的各个实施例均采用递进的方式描述，各个实施例之间相同相似的部分互相参见即可，每个实施例重点说明的都是与其他实施例的不同之处。尤其，对于系统实施例而言，由于其基本相似于方法实施例，所以描述的比较简单，相关之处参见方法实施例的部分说明即可。

以上所述仅为本申请的实施例而已，并不用于限制本申请。对于本领域技术人员来说，本申请可以有各种更改和变化。凡在本申请的精神和原理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权利要求范围之内。

权 利 要 求 书

1、一种模型的训练方法，其特征在于，所述方法包括:

获取多个用户数据对，其中，所述每个用户数据对中的两组用户数据的数据字段有相同的部分;

获取每个用户数据对所对应的用户相似度，所述用户相似度为每个用户数据对中的两组用户数据对应的用户之间的相似度;

根据所述每个用户数据对所对应的用户相似度和所述多个用户数据对，确定用于训练预设的分类模型的样本数据;

基于所述样本数据对所述分类模型进行训练，得到相似度分类模型。

2、根据权利要求 1 所述的方法，其特征在于，所述获取每个用户数据对所对应的用户相似度，包括:

获取第一用户数据对所对应的用户的生物特征，其中，所述第一用户数据对为所述多个用户数据对中的任意用户数据对;

根据所述第一用户数据对所对应的用户的生物特征，确定所述第一用户数据对所对应的用户相似度。

3、根据权利要求 2 所述的方法，其特征在于，所述生物特征包括面部图像特征，

所述获取第一用户数据对所对应的用户的生物特征，包括:

获取第一用户数据对所对应的用户的面部图像;

对所述面部图像进行特征提取，得到面部图像特征;

相应的，所述根据所述第一用户数据对所对应的用户的生物特征，确定所述第一用户数据对所对应的用户相似度，包括:

根据所述第一用户数据对所对应的用户的面部图像特征，确定所述第一用户数据对所对应的用户相似度。

4、根据权利要求 2 所述的方法，其特征在于，所述生物特征包括语音特征，

所述获取第一用户数据对所对应的用户的生物特征，包括:

获取第一用户数据对所对应的用户的语音数据;

对所述语音数据进行特征提取，得到语音特征;

相应的，所述根据所述第一用户数据对所对应的用户的生物特征，确定所述第一用户数据对所对应的用户相似度，包括:

根据所述第一用户数据对所对应的用户的语音特征，确定所述第一用户数据对所对应的用户相似度。

5、根据权利要求 1 所述的方法，其特征在于，所述根据所述每个用户数据对所对应的用户相似度和所述多个用户数据对，确定用于训练分类模型的样本数据，包括:

对所述多个用户数据对中的每个用户数据对进行特征提取，得到每个用户数据对中的两组用户数据之间相关联的用户特征;

根据所述每个用户数据对中用户数据之间相关联的用户特征和所述每个用户数据对所对应的用户相似度，确定用于训练分类模型的样本数据。

6、根据权利要求 5 所述的方法，其特征在于，所述根据所述每个用户数据对中的两组用户数据之间相关联的用户特征和所述每个用户数据对所对应的用户相似度，确定用于训练分类模型的样本数据，包括:

根据每个用户数据对所对应的用户相似度和预定的相似度阈值，从所述多个用户数据对所对应的用户特征中选取正样本特征和负样本特征;

将所述正样本特征和负样本特征作为用于训练分类模型的样本数据。

7、根据权利要求 6 所述的方法，其特征在于，所述用户特征包括户籍维度特征、姓名维度特征、社交特征和兴趣爱好特征; 所述户籍维度特征包括用户身份信息的特征，所述姓名维度特征包括用户姓名信息的特征和用户姓氏的稀缺程度的特征，所述社交特征包括用户的社会关系信息的特征。

8、根据权利要求 6 所述的方法，其特征在于，所述正样本特征和负样本特征中包含的特征数目相同。

9、根据权利要求 1-8 中任一项所述的方法，其特征在于，所述相似度分类模型为二分类器模型。

10、一种数据相似度的确定方法，其特征在于，所述方法包括:

获取待测用户数据对;

对所述待测用户数据对中每组待测用户数据进行特征提取，得到待测用户特征;

根据所述待测用户特征和预先训练的相似度分类模型，确定所述待测用户数据对中的两组待测用户数据对应的用户之间的相似度。

11、根据权利要求 10 所述的方法，其特征在于，所述方法还包括:

如果所述待测用户数据对中的两组待测用户数据对应的用户之间的相似度大于预定相似度分类阈值，则确定所述待测用户数据对所对应的待测用户为双胞胎。

12、一种模型的训练装置，其特征在于，所述装置包括:

数据获取模块，用于获取多个用户数据对，其中，所述每个用户数据对中的两组用户数据的数据字段有相同的部分;

相似度获取模块，用于获取每个用户数据对所对应的用户相似度，所述用户相似度为每个用户数据对中的两组用户数据对应的用户之间的相似度;

样本数据确定模块，用于根据所述每个用户数据对所对应的用户相似度和所述多个用户数据对，确定用于训练预设的分类模型的样本数据;

模型训练模块，用于基于所述样本数据对所述分类模型进行训练，得到相似度分类模型。

13、根据权利要求 12 所述的装置，其特征在于，所述相似度获取模块，包括:

生物特征获取单元，用于获取第一用户数据对所对应的用户的生物特征，其中，所述第一用户数据对为所述多个用户数据对中的任意用户数据对;

相似度获取单元，用于根据所述第一用户数据对所对应的用户的生物特征，确定所述第一用户数据对所对应的用户相似度。

14、根据权利要求 13 所述的装置，其特征在于，所述生物特征包括面部

图像特征，

所述生物特征获取单元，用于获取第一用户数据对所对应的用户的面部图像；对所述面部图像进行特征提取，得到面部图像特征；

相应的，所述相似度获取单元，用于根据所述第一用户数据对所对应的用户的面部图像特征，确定所述第一用户数据对所对应的用户相似度。

15、根据权利要求 13 所述的装置，其特征在于，所述生物特征包括语音特征，

所述生物特征获取单元，用于获取第一用户数据对所对应的用户的语音数据；对所述语音数据进行特征提取，得到语音特征；

相应的，所述相似度获取单元，用于根据所述第一用户数据对所对应的用户的语音特征，确定所述第一用户数据对所对应的用户相似度。

16、根据权利要求 12 所述的装置，其特征在于，所述样本数据确定模块，包括：

特征提取单元，用于对所述多个用户数据对中的每个用户数据对进行特征提取，得到每个用户数据对中的两组用户数据之间相关联的用户特征；

样本数据确定单元，用于根据所述每个用户数据对中的两组用户数据之间相关联的用户特征和所述每个用户数据对所对应的用户相似度，确定用于训练分类模型的样本数据。

17、根据权利要求 16 所述的装置，其特征在于，所述样本数据确定单元，用于根据每个用户数据对所对应的用户相似度和预定的相似度阈值，从所述多个用户数据对所对应的用户特征中选取正样本特征和负样本特征；将所述正样本特征和负样本特征作为用于训练分类模型的样本数据。

18、根据权利要求 17 所述的装置，其特征在于，所述用户特征包括户籍维度特征、姓名维度特征、社交特征和兴趣爱好特征；所述户籍维度特征包括用户身份信息的特征，所述姓名维度特征包括用户姓名信息的特征和用户姓氏的稀缺程度的特征，所述社交特征包括用户的社会关系信息的特征。

19、根据权利要求 17 所述的装置，其特征在于，所述正样本特征和负样本特征中包含的特征数目相同。

20、根据权利要求 12-19 中任一项所述的装置，其特征在于，所述相似度分类模型为二分类器模型。

21、一种数据相似度的确定装置，其特征在于，所述装置包括:

待测数据获取模块，用于获取待测用户数据对;

特征提取模块，用于对所述待测用户数据对中每组待测用户数据进行特征提取，得到待测用户特征;

相似度确定模块，用于根据所述待测用户特征和预先训练的相似度分类模型，确定所述待测用户数据对中的两组待测用户数据对应的用户之间的相似度。

22、根据权利要求 21 所述的装置，其特征在于，所述装置还包括:

相似度分类模块，用于如果所述待测用户数据对中的两组待测用户数据对应的用户之间的相似度大于预定相似度分类阈值，则确定所述待测用户数据对所对应的待测用户为双胞胎。

23、一种模型的训练设备，所述设备包括:

处理器；以及

被安排成存储计算机可执行指令的存储器，所述可执行指令在被执行时使所述处理器执行以下操作:

获取多个用户数据对，其中，所述每个用户数据对中的两组用户数据的数据字段有相同的部分;

获取每个用户数据对所对应的用户相似度，所述用户相似度为每个用户数据对中的两组用户数据对应的用户之间的相似度;

根据所述每个用户数据对所对应的用户相似度和所述多个用户数据对，确定用于训练预设的分类模型的样本数据;

基于所述样本数据对所述分类模型进行训练，得到相似度分类模型。

24、一种数据相似度的确定设备，所述设备包括:

处理器；以及

被安排成存储计算机可执行指令的存储器，所述可执行指令在被执行时使所述处理器执行以下操作：

获取待测用户数据对；

对所述待测用户数据对中每组待测用户数据进行特征提取，得到待测用户特征；

根据所述待测用户特征和预先训练的相似度分类模型，确定所述待测用户数据对中的两组待测用户数据对应的用户之间的相似度。

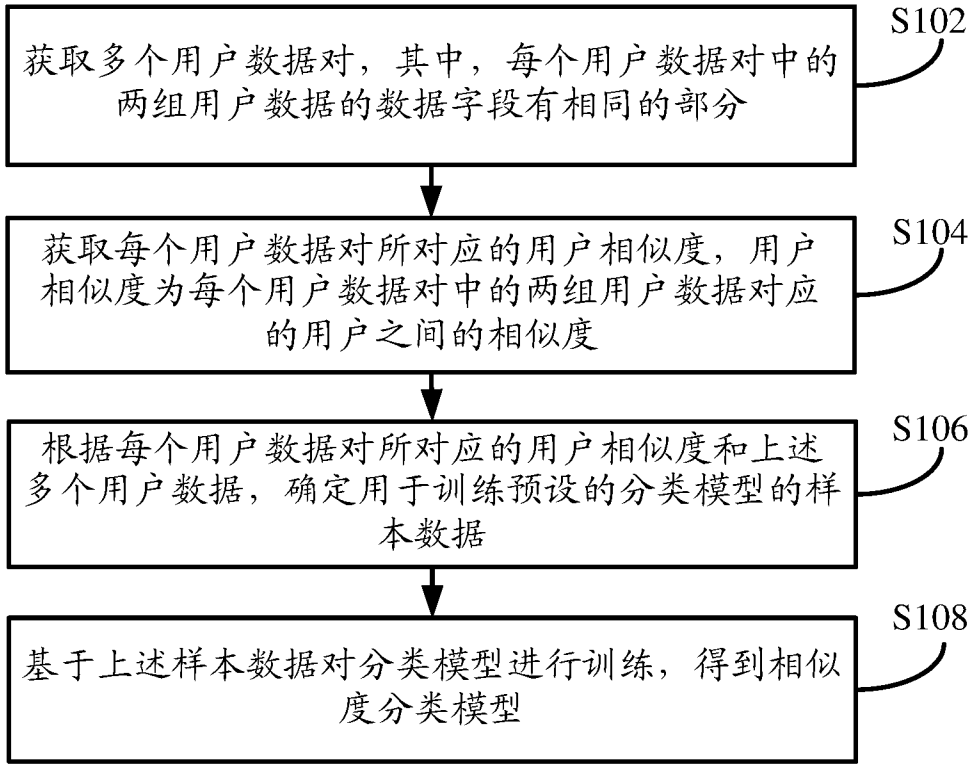
获取多个用户数据对，其中，每个用户数据对中的两组用户数据的数据字段有相同的部分
S102
获取每个用户数据对所对应的用户相似度，用户相似度为每个用户数据对中的两组用户数据对应的用户之间的相似度
S104
根据每个用户数据对所对应的用户相似度和上述多个用户数据，确定用于训练预设的分类模型的样本数据
S106
基于上述样本数据对分类模型进行训练，得到相似度分类模型
S108

图 1

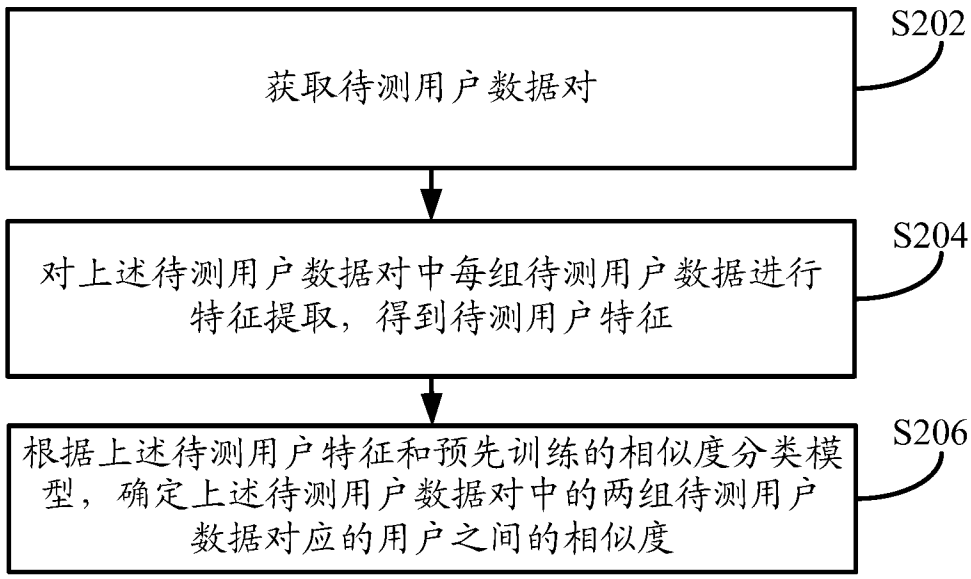
获取待测用户数据对
S202
对上述待测用户数据对中每组待测用户数据进行特征提取，得到待测用户特征
S204
根据上述待测用户特征和预先训练的相似度分类模型，确定上述待测用户数据对中的两组待测用户数据对应的用户之间的相似度
S206

图 2

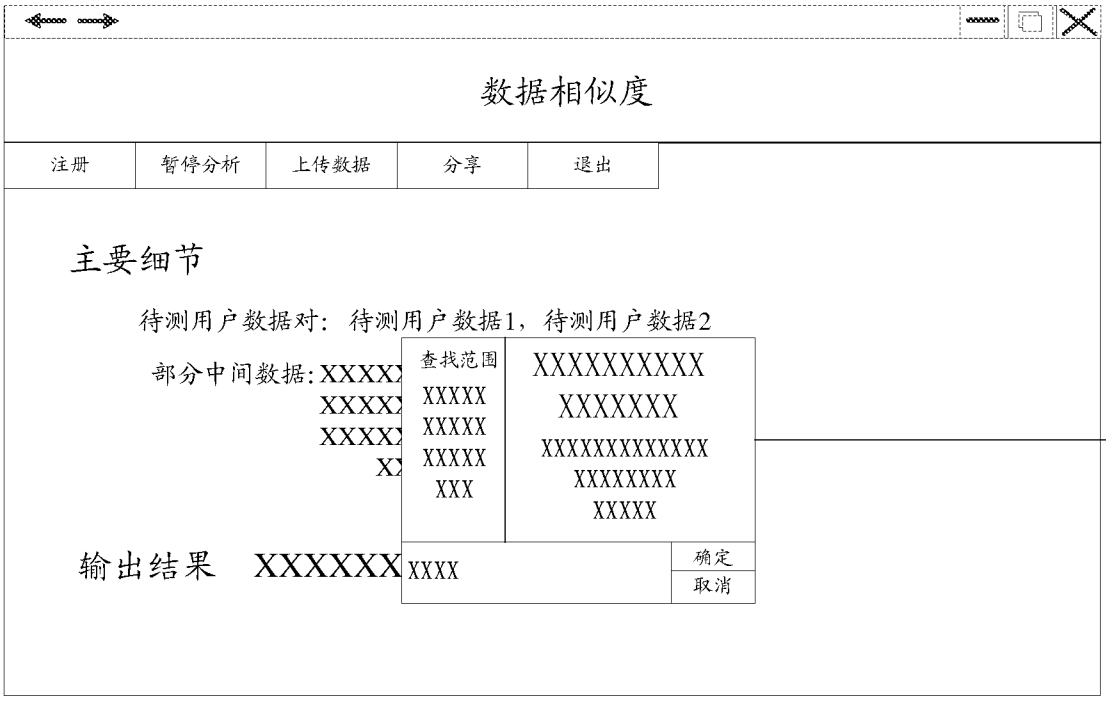
数据相似度
注册
暂停分析
上传数据
分享
退出
主要细节
待测用户数据对：待测用户数据1，待测用户数据2
部分中间数据:XXXX
XXXX
XXXX
XX
查找范围
XXXXX
XXXXX
XXXXX
XXX
XXXXXXXXXX
XXXXXXX
XXXXXXXXXXXXX
XXXXXXXX
XXXXX
输出结果 XXXXXX
XXXX
确定
取消
提示框

图 3

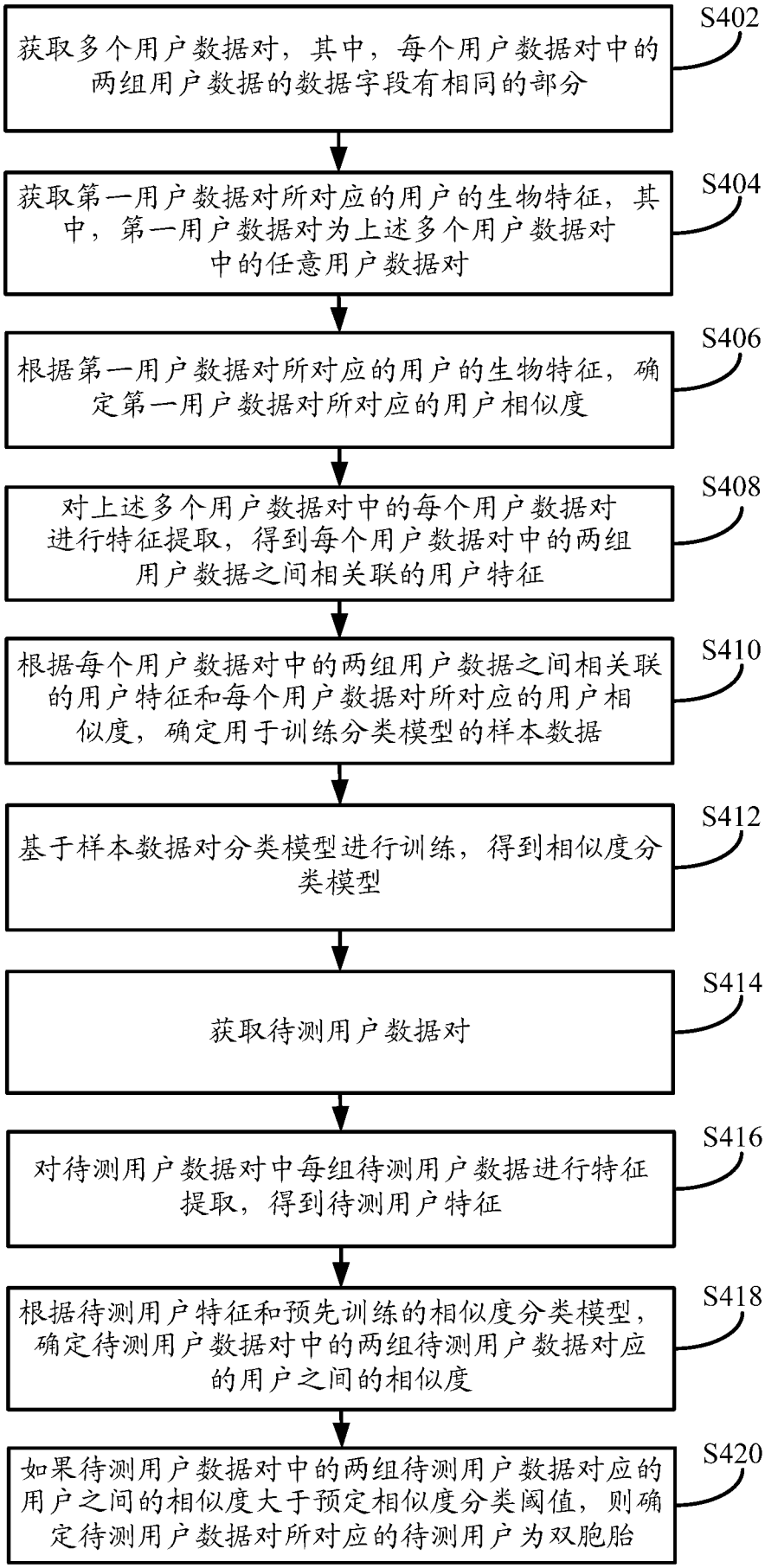
S402
获取多个用户数据对，其中，每个用户数据对中的两组用户数据的数据字段有相同的部分
S404
获取第一用户数据对所对应的用户的生物特征，其中，第一用户数据对为上述多个用户数据对中的任意用户数据对
S406
根据第一用户数据对所对应的用户的生物特征，确定第一用户数据对所对应的用户相似度
S408
对上述多个用户数据对中的每个用户数据对进行特征提取，得到每个用户数据对中的两组用户数据之间相关联的用户特征
S410
根据每个用户数据对中的两组用户数据之间相关联的用户特征和每个用户数据对所对应的用户相似度，确定用于训练分类模型的样本数据
S412
基于样本数据对分类模型进行训练，得到相似度分类模型
S414
获取待测用户数据对
S416
对待测用户数据对中每组待测用户数据进行特征提取，得到待测用户特征
S418
根据待测用户特征和预先训练的相似度分类模型，确定待测用户数据对中的两组待测用户数据对应的用户之间的相似度
S420
如果待测用户数据对中的两组待测用户数据对应的用户之间的相似度大于预定相似度分类阈值，则确定待测用户数据对所对应的待测用户为双胞胎

图 4

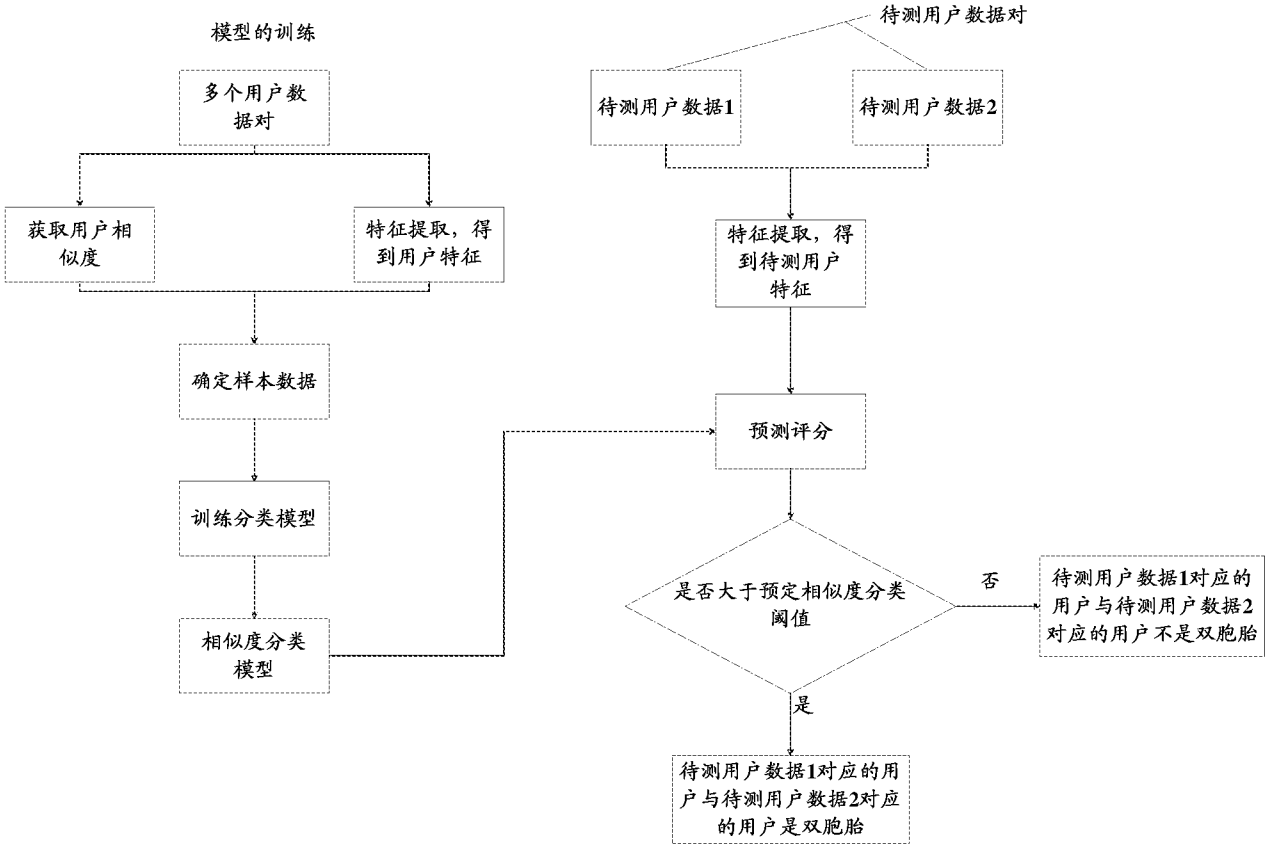
模型的训练
多个用户数据对
获取用户相似度
特征提取，得到用户特征
确定样本数据
训练分类模型
相似度分类模型
待测用户数据对
待测用户数据1
待测用户数据2
特征提取，得到待测用户特征
预测评分
是否大于预定相似度分类阈值
否
待测用户数据1对应的用户与待测用户数据2对应的用户不是双胞胎
是
待测用户数据1对应的用户与待测用户数据2对应的用户是双胞胎

图 5

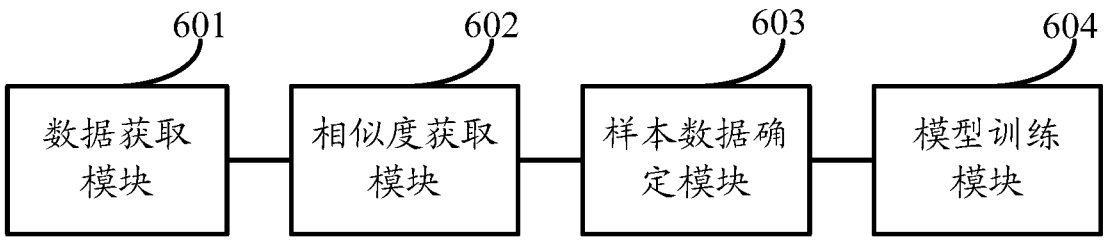
601
602
603
604
数据获取模块
相似度获取模块
样本数据确定模块
模型训练模块

图 6

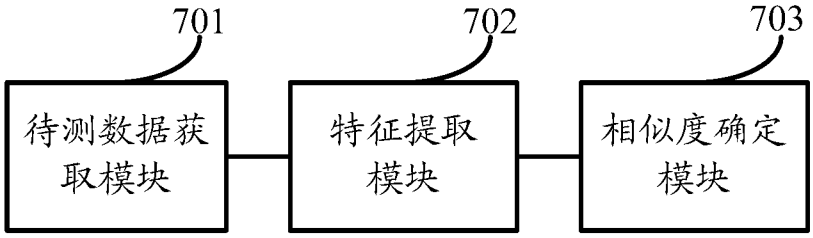
701
702
703
待测数据获取模块
特征提取模块
相似度确定模块

图 7

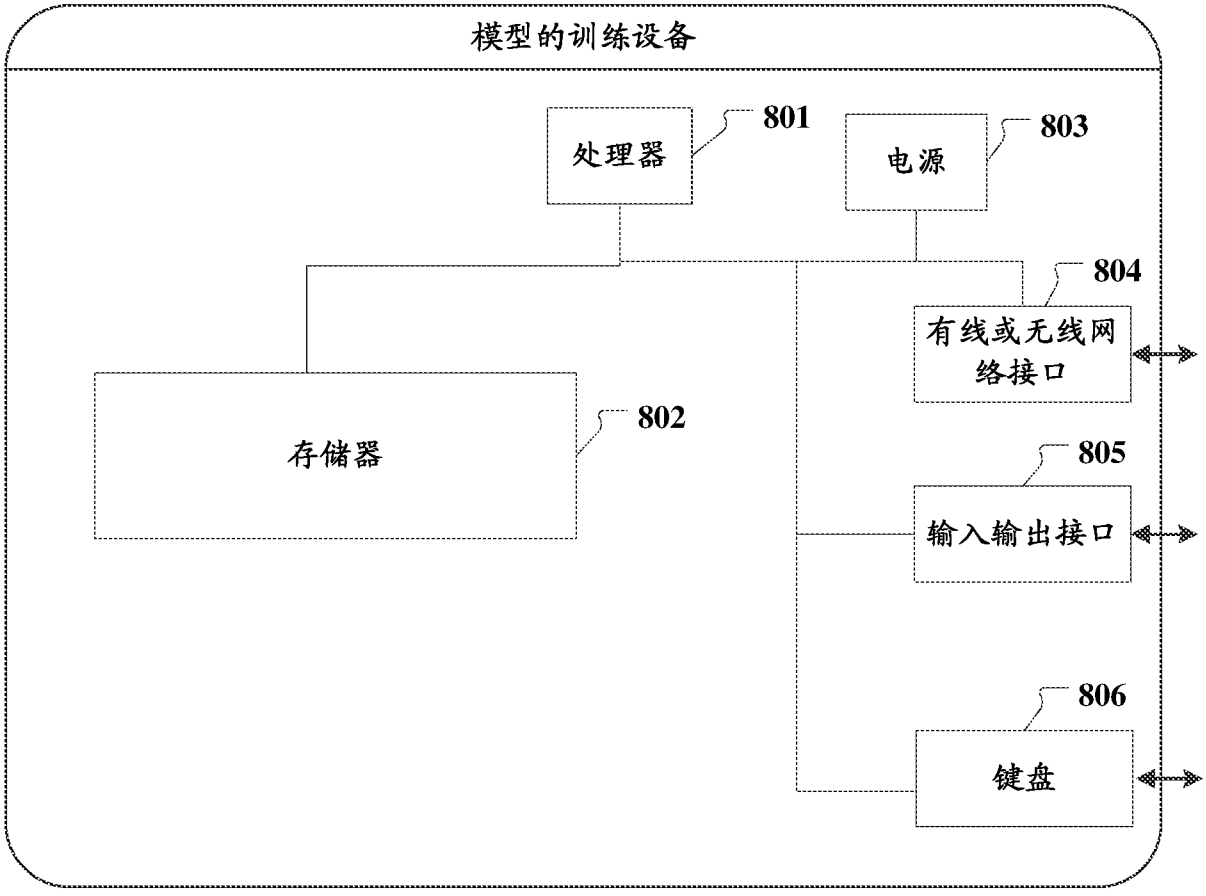
模型的训练设备
处理器
801
电源
803
804
有线或无线网络接口
存储器
802
805
输入输出接口
806
键盘

图 8

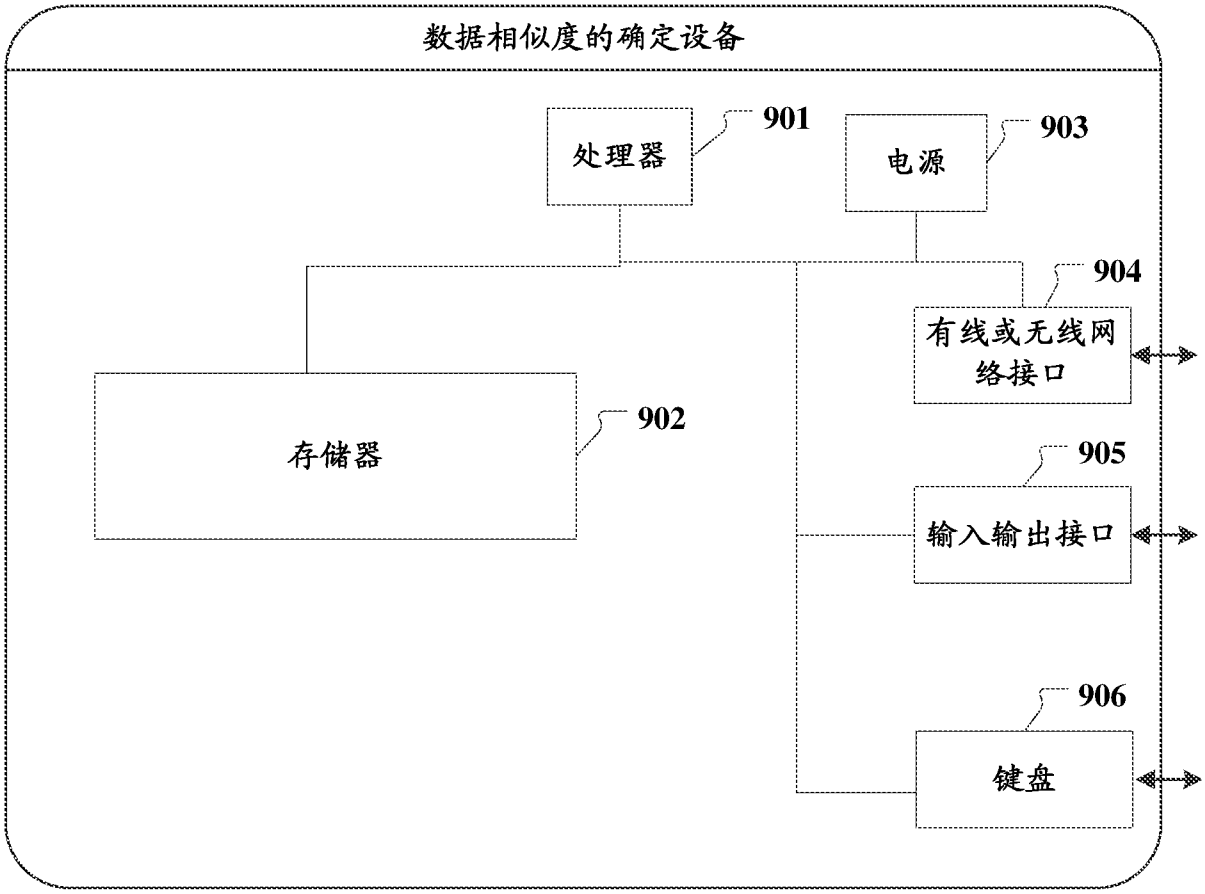
数据相似度的确定设备
处理器
901
电源
903
904
有线或无线网络接口
存储器
902
905
输入输出接口
906
键盘

图 9

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2018/096252

A. CLASSIFICATION OF SUBJECT MATTER

G06K 9/00(2006.01)i; G06K 9/62(2006.01)n

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)

G06K 9/-

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)

CNABS, CNTXT, CNKI, DWPI, SIPOABS: 相似, 模型, 样本, 训练, 分类, 人脸, 面部, 生物, similarity, model, sample, train +, classification, face, biometric

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	CN 103745242 A (INSTITUTE OF AUTOMATION, CHINESE ACADEMY OF SCIENCES) 23 April 2014 (2014-04-23) description, paragraphs [0022]-[0063]	1-24
A	CN 102129574 A (BEIJING VIMICRO CORP.) 20 July 2011 (2011-07-20) entire document	1-24
A	CN 105488463 A (KONKA GROUP CO., LTD.) 13 April 2016 (2016-04-13) entire document	1-24
PX	CN 107609461 A (ALIBABA GROUP HOLDING LIMITED) 19 January 2018 (2018-01-19) claims 1-24	1-24

☐ 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	Date of mailing of the international search report
28 September 2018	**22 October 2018**

Name and mailing address of the ISA/CN	Authorized officer
State Intellectual Property Office of the P. R. China **No. 6, Xitucheng Road, Jimenqiao Haidian District, Beijing** **100088** **China**	
Facsimile No. **(86-10)62019451**	Telephone No.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2018/096252

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)			Publication date (day/month/year)
CN	103745242	A	23 April 2014	None			
CN	102129574	A	20 July 2011	CN	102129574	B	07 December 2016
CN	105488463	A	13 April 2016	None			
CN	107609461	A	19 January 2018	None			

国际检索报告

国际申请号 PCT/CN2018/096252

A. 主题的分类

G06K 9/00(2006.01)i；G06K 9/62(2006.01)n

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)

G06K 9/-

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）)

CNABS, CNTXT, CNKI, DWPI, SIPOABS：相似，模型，样本，训练，分类，人脸，面部，生物，similarity, model, sample, train+, classification, face, biometric

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
X	CN 103745242 A (中国科学院自动化研究所) 2014年 4月 23日 (2014-04-23) 说明书第[0022]-[0063]段	1-24
A	CN 102129574 A (北京中星微电子有限公司) 2011年 7月 20日 (2011-07-20) 全文	1-24
A	CN 105488463 A (康佳集团股份有限公司) 2016年 4月 13日 (2016-04-13) 全文	1-24
PX	CN 107609461 A (阿里巴巴集团控股有限公司) 2018年 1月 19日 (2018-01-19) 权利要求1-24	1-24

☐ 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	国际检索报告邮寄日期
2018年 9月 28日	2018年 10月 22日

ISA/CN的名称和邮寄地址	受权官员
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	朱晓莉
传真号 (86-10)62019451	电话号码 (86-10)62411672

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2018/096252

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	103745242	A	2014年 4月 23日	无			
CN	102129574	A	2011年 7月 20日	CN	102129574	B	2016年 12月 7日
CN	105488463	A	2016年 4月 13日	无			
CN	107609461	A	2018年 1月 19日	无			