(12) 按照专利合作条约所公布的国际申请

(19) 世界知识产权组织
国 际 局

(43) 国际公布日
2017 年 6 月 29 日 (29.06.2017)

WIPO | PCT

(10) 国际公布号
WO 2017/107422 A1

(51) 国际专利分类号:
*G06K 9/62* (2006.01)

(21) 国际申请号: PCT/CN2016/087685

(22) 国际申请日: 2016 年 6 月 29 日 (29.06.2016)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
201510965231.0 2015 年 12 月 21 日 (21.12.2015) CN

(71) 申请人: **百度在线网络技术（北京）有限公司 (BAIDU ONLINE NETWORK TECHNOLOGY (BEIJING) CO., LTD.)** [CN/CN]; 中国北京市海淀区上地十街 10 号百度大厦三层, Beijing 100085 (CN)。

(72) 发明人: **马守玉 (MA, Shouyu)**; 中国北京市海淀区上地十街 10 号百度大厦三层, Beijing 100085 (CN)。
**徐锐 (XU, Rui)**; 中国北京市海淀区上地十街 10 号百度大厦三层, Beijing 100085 (CN)。

(74) 代理人: **北京英赛嘉华知识产权代理有限责任公司 (INSIGHT INTELLECTUAL PROPERTY LIMITED)**; 中国北京市海淀区知春路甲 48 号盈都大厦 A 座 19A, Beijing 100098 (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, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JP, KE, KG, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW。

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

[见续页]

(54) **Title**: METHOD AND DEVICE FOR USER GENDER IDENTIFICATION

(54) **发明名称** : 一种用户性别识别方法及装置

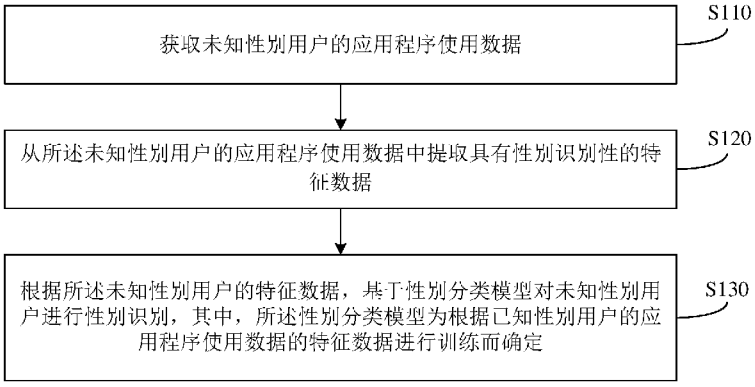



图 1

S110 ACQUIRE APPLICATION USE DATA OF A GENDER UNKNOWN USER
S120 EXTRACT FEATURE DATA HAVING GENDER IDENTIFIABILITY FROM THE APPLICATION USE DATA OF THE GENDER UNKNOWN USER
S130 IDENTIFY THE GENDER OF THE GENDER UNKNOWN USER BASED ON A GENDER DISTINGUISHING MODEL ON THE BASIS OF THE FEATURE DATA OF THE GENDER UNKNOWN USER, WHERE THE GENDER DISTINGUISHING MODEL IS DETERMINED BY TRAINING ON THE BASIS OF FEATURE DATA OF APPLICATION USE DATA OF GENDER KNOWN USERS

(57) **Abstract**: A method for user gender identification, comprising: acquiring application use data of a gender unknown user (S110 and S240), extracting feature data having gender identifiability from the application use data of the gender unknown user (S120 and S250), and identifying the gender of the gender unknown user based on a gender distinguishing model on the basis of the feature data of the gender unknown user, where the gender distinguishing model is determined by training on the basis of feature data of application use data of gender known users (S130 and S260). Employment of the method solves the technical problem in the prior art of possible involvement of user privacy or expensive costs in user gender identification.

(57) **摘要**:

[见续页]

WO 2017/107422 A1

**本国际公布**:

— 包括国际检索报告(条约第 21 条(3))。

一种用户性别识别方法，其中，包括：获取未知性别用户的应用程序使用数据（S110,S240），从所述未知性别用户的应用程序使用数据中提取具有性别识别性的特征数据（S120,S250），根据所述未知性别用户的特征数据，基于性别分类模型对未知性别用户进行性别识别，其中，所述性别分类模型为根据已知性别用户的应用程序使用数据的特征数据进行训练而确定（S130,S260）。采用该方法，可以解决现有技术中用户性别识别时，可能涉及用户隐私或者成本较高的技术问题。

# 一种用户性别识别方法及装置

## 技术领域

本发明涉及数据处理技术领域，尤其涉及一种用户性别识别方法及装置。

## 背景技术

随着移动设备的普及和移动宽带网络的建设，当前我们已进入移动互联网的新时代。移动应用分发作为移动互联网的入口之一，也决定了各个厂商在移动互联网上掌控能力的强弱。因此，越来越多的厂商向诸如手机、平板电脑、笔记本、台式机之类的终端设备提供了各种应用程序。为了个性化地给用户推荐应用程序，我们往往需要知道用户性别。

现有技术中用户性别识别主要基于以下几种方法：一种是根据用户注册时填写的资料直接识别；另一种是依据人脸识别技术，通过扫描用户面部特征，识别用户性别。

但是，通过用户填写注册信息或者人脸识别技术，可能涉及用户隐私；而面部识别技术，成本过高。

## 发明内容

有鉴于此，本发明实施例提供一种用户性别识别方法及装置，以解决现有技术中用户性别识别时，可能涉及用户隐私或者成本较高的技术问题。

第一方面，本发明实施例提供了一种用户性别识别方法，所述方法包括：

获取未知性别用户的应用程序使用数据；

从所述未知性别用户的应用程序使用数据中提取具有性别识别性的特征数据；

根据所述未知性别用户的特征数据，基于性别分类模型对未知性别用户进行性别识别，其中，所述性别分类模型为根据已知性别用户的应用程序使用数据的特征数据进行训练而确定。

第二方面，本发明实施例还提供了一种用户性别识别装置，所述装置包括：

第一使用数据获取模块，用于获取未知性别用户的应用程序使用数据；

第一特征数据提取模块，用于从所述未知性别用户的应用程序使用数据中提取具有性别识别性的特征数据；

性别识别模块，用于根据所述未知性别用户的特征数据，基于性别分类模型对未知性别用户进行性别识别，其中，所述性别分类模型为根据已知性别用户的应用程序使用数据的特征数据进行训练而确定。

本发明实施例提供的用户性别识别方法和装置，通过获取未知性别用户的应用程序使用数据，从中提取具有性别识别性的特征数据，并应用性别分类模型学习所述特征数据，进而判断用户性别，实现了仅通过用户应用程序使用数据识别用户性别的目的。由于用户应用程序使用数据比较丰富，因此性别识别过程中准确率较高，并且可以解决现有技术中用户性别识别时，可能涉及用户隐私或者成本较高的技术问题。

## 附图说明

通过阅读参照以下附图所作的对非限制性实施例所作的详细描述，本发明的其它特征、目的和优点将会变得更明显：

图 1 为本发明实施例一提供的一种用户性别识别方法的流程示意图；

图 2 为本发明实施例二提供的一种用户性别识别方法的流程示意图；

图 3 为本发明实施例三提供的一种用户性别识别装置的结构框图；

图 4 示出了适于用来实现本申请实施例的终端设备或服务器的计算机系统的结构示意图。

## 具体实施方式

下面结合附图和实施例对本发明作进一步的详细说明。可以理解的是，此处所描述的具体实施例仅仅用于解释本发明，而非对本发明的限定。另外还需要说明的是，为了便于描述，附图中仅示出了与本发明相关的部分而非全部内容。

实施例一

图 1 为本发明实施例一提供的一种用户性别识别方法的流程示意图。本实施例可适用于识别未知性别用户的性别的情况。该方法可以由用户性别识别装置来执行，其中该装置可以由软件和/或硬件来实现。如图 1 所示，本实施例提供的用户性别识别方法具体可以包括如下：

S110、获取未知性别用户的应用程序使用数据。

示例性的，获取未知性别用户的应用程序使用数据，该应用程序使用数据可以包括应用程序安装列表和应用程序使用行为中的至少一个。应用程序安装列表即用户的移动终端中安装的各种应用程序的一个集合，应用程序使用行为即用户对应用程序具体的操作行为，例如对应用程序的使用次数，或者使用频率，或者使用时间等。

S120、从所述未知性别用户的应用程序使用数据中提取具有性别识别性的特征数据。

示例性的，具有性别识别性的特征数据，即帮助识别未知性别用户的性别的特征数据，或者具有性别属性的特征数据。从未知性别用户的应用程序安装列表和/或应用程序使用行为中提取具有性别识别性的特征

数据。

S130、根据所述未知性别用户的特征数据，基于性别分类模型对未知性别用户进行性别识别，其中，所述性别分类模型为根据已知性别用户的应用程序使用数据的特征数据进行训练而确定。

示例性的，基于性别分类模型，通过未知性别用户的特征数据，对未知性别用户进行性别识别，识别所述未知性别用于性别为男或者为女，以及对应的概率，其中，性别分类模型可以为根据已知性别用户的应用程序使用数据进行训练而确定的。

具体的，根据未知性别用户的特征数据，基于性别分类模型对未知性别用户进行性别识别，例如：具体识别结果可以是未知性别用户为男性的概率为 80%，为女性的概率为 20%。进一步的，当未知性别用户的特征数据发生改变时，更新所述概率。具体可以是未知性别用户的应用程序安装列表发生改变，例如用户新增或者删除某些应用程序时，更新所述概率，还可以是未知性别用户的应用程序使用行为发生改变时，更新所述概率。

本发明实施例一提供的用户性别识别方法，通过获取未知性别用户的应用程序使用数据，从中提取具有性别识别性的特征数据，并应用性别分类模型学习所述特征数据，进而判断用户性别，实现了仅通过用户应用程序使用数据识别用户性别的目的。由于用户应用程序使用数据比较丰富，因此性别识别过程中准确率较高，并且可以解决现有技术中用户性别识别时，可能涉及用户隐私或者成本较高的问题。

实施例二

本实施例以上述实施例为基础，提供一种用户性别识别方法。图 2 为本发明实施例二提供的一种用户性别识别方法的流程示意图。如图 2 所示，本实施例提供的用户性别识别方法可以包括如下：

S210、获取已知性别用户的应用程序使用数据。

示例性的，应用程序使用数据可以包括应用程序安装列表和应用程序使用行为中的至少一个。这里，由于已知性别用户的应用程序安装列表可能涉及用户的隐私，出于保护隐私的目的，部分已知性别用户的应用程序安装列表可能是加密存储的，这是将加密存储的应用程序安装列表进行必要的数据格式转换，通过转换解密为相应的已知性别用户的应用程序安装列表。

S220、从所述已知性别用户的应用程序使用数据中提取具有性别识别性的特征数据。

示例性的，从已知性别用户的应用程序使用数据中提取具有性别识别性的特征数据，可以包括下述至少一项:

根据所述应用程序安装列表，提取具有性别识别性的第一特征数据，所述第一特征数据为至少两个设定类别的应用程序的数量；

根据所述应用程序安装列表，提取具有性别识别性的第二特征数据，所述第二特征数据为至少两个应用程序安装人数之间的性别比例；

根据所述应用程序使用行为，提取具有性别识别性的第三特征数据，所述第三特征数据为应用程序中同类使用行为的统计数据。

具体的，根据所述应用程序安装列表，提取具有性别识别性的第一特征数据，所述第一特征数据为至少两个设定类别的应用程序的数量，即根据已知性别用户的应用程序安装列表，提取至少两个设定类别的应用程序的数量。由于应用程序在安装使用的时候已经规定了应用程序的类别，例如百度地图的类别为导航，百度糯米的类别为生活。这里，根据已知性别用户的应用程序安装列表，可以提取诸如导航、生活、娱乐、购物、社交、财务、天气、游戏等各个类别的应用程序的数量，将上述各个类别中至少两个设定类别的应用程序的数量作为第一特征数据。

具体的，根据所述应用程序安装列表，提取具有性别识别性的第二

特征数据，所述第二特征数据为至少两个应用程序安装人数之间的性别比例，即根据已知性别用户的应用程序安装列表，提取至少两个应用程序安装人数之间的性别比例。这里，根据大量的已知性别用户的应用程序安装列表，可以提取诸如体育头条、美颜相机、美妆相机等应用程序安装人数之间的性别比例，将上述各个应用程序中至少两个应用程序安装人数之间的性别比例作为第二特征数据。这里，由于部分应用程序可能存在安装人数较少，或者某一性别用户安装人数为零的情况，这样计算出来的应用程序安装人数之间的比例就会存在过小或者过大的问题，此时可以利用拉普拉斯平滑方法，将应用程序的初始安装人数默认值由原先的 0 改为 1。进一步的，当应用程序安装人数之间的性别比例数值过大时，还可以对所述数值进行进一步运算，例如取对数运算，以方便接来下的训练性别分类模型。

具体的，根据所述应用程序使用行为，提取具有性别识别性的第三特征数据，所述第三特征数据为应用程序中同类使用行为的统计数据，即根据已知性别用户的应用程序使用行为，提取应用程序中同类使用行为的统计数据。具体可以是将已知性别用户的各种使用行为进行聚类，形成同类使用行为，然后提取同类使用行为的统计数据。可选的，应用程序中同类使用行为的统计数据可以包括下述至少一项：消费应用程序的消费行为中对消费金额、消费频率和/或消费品分类的统计数量或比例；阅读器应用程序的阅读行为中对阅读内容分类的统计数量或比例；健身应用程序的健身训练行为中，对健身训练内容、健身训练强度和/或健身训练频率的统计数量或比例，将上述同类使用行为的统计数据中的至少一项作为第三特征数据。

S230、根据所述特征数据，训练性别分类模型。

示例性的，根据该特征数据，使用有监督的机器学习方法训练性别分类模型，具体可以是：

根据第一特征数据，基于迭代决策树学习算法和算法参数，生成已知性别用户配置信息，训练迭代决策树模型。

具体的，迭代决策树模型作为一种集成学习算法在特征维度较低时，训练效率高，且准确率较高，这里，根据至少两个设定类别的应用程序的数量，训练迭代决策树模型。

根据所述第一特征数据以及第二特征数据，基于稀疏逻辑回归学习算法和算法参数，生成已知性别用户稀疏逻辑回归配置信息，训练稀疏逻辑回归模型。具体的，由于第二特征数据比较稀疏，常规的逻辑回归可能导致模型过拟合，因此这里采用带 L1 惩罚项的稀疏逻辑回归模型，基于稀疏逻辑回归学习算法和算法参数,训练稀疏逻辑回归模型。示例性的，通过上述的模型训练，可以得到不同性别用户的模型特征，例如，男性的购物类别应用程序为 1-2 个，女性的购物类别应用程序为 5-6 个；男性的游戏类别应用程序为 5-6 个，女性的游戏类别的应用程序为 1-2 个。还可以得到典型应用程序安装人数之间的性别比例，例如，体育头条的安装人数男女之间的性别比例为 8：1，美颜相机安装人数男女之间的性别比例为 1：10。

进一步的，上述模型的特征可以单独用于性别识别，也可以结合用于性别识别。

S240、获取未知性别用户的应用程序使用数据。

S250、从所述未知性别用户的应用程序使用数据中提取具有性别识别性的特征数据。

S260、根据所述未知性别用户的特征数据，基于性别分类模型对未知性别用户进行性别识别，其中，所述性别分类模型为根据已知性别用户的应用程序使用数据的特征数据进行训练而确定。

本发明实施例二提供的用户性别识别方法，通过获取已知性别用户的应用程序使用数据，并从中提取具有性别识别性的特征数据，然后根

据该特征数据，训练性别模型，最后通过性别分类模型识别未知性别用户的性别。采用该方法，从已知性别用户的应用程序使用数据中学习有效的性别分类模型，通过行别分类模型识别未知性别用户的性别，性别识别准确率更高。

实施例三

图 3 为本发明实施例三提供的一种用户性别识别装置的结构框图，该装置可由软件和/或硬件实现。本实施例可适用于想要识别未知性别用户的性别的情况，如图 3 所示，该装置包括：第一使用数据获取模块 310 第一特征数据提取模块 320 以及性别识别模块 330。

其中，第一使用数据获取模块 310，用于获取未知性别用户的应用程序使用数据；第一特征数据提取模块 320，用于从所述未知性别用户的应用程序使用数据中提取具有性别识别性的特征数据；性别识别模块 330，用于根据所述未知性别用户的特征数据，基于性别分类模型对未知性别用户进行性别识别，其中，所述性别分类模型为根据已知性别用户的应用程序使用数据的特征数据进行训练而确定。

进一步的，所述应用程序使用数据可以包括：

应用程序安装列表和应用程序使用行为中的至少一个。

进一步的，所述装置还可以包括：

第二使用数据获取模块，用于获取已知性别用户的应用程序使用数据；

第二特征数据提取模块，用于从所述已知性别用户的应用程序使用数据中提取具有性别识别性的特征数据；

模型训练模块，用于根据所述特征数据，训练性别分类模型。

进一步的，第二特征数据提取模块具体用于执行下述至少一项:

根据所述应用程序安装列表，提取具有性别识别性的第一特征数据，

所述第一特征数据为至少两个设定类别的应用程序的数量；

根据所述应用程序安装列表，提取具有性别识别性的第二特征数据，所述第二特征数据为至少两个应用程序安装人数之间的性别比例；

根据所述应用程序使用行为，提取具有性别识别性的第三特征数据，所述第三特征数据为应用程序中同类使用行为的统计数据。

进一步的，所述模型训练模块，可以包括：

迭代决策树模型训练子模块，用于根据所述第一特征数据，基于迭代决策树学习算法和算法参数，生成已知性别用户迭代决策树配置信息，训练迭代决策树模型；

稀疏逻辑回归模型训练子模块，用于根据所述第一特征数据以及第二特征数据，基于稀疏逻辑回归学习算法和算法参数，生成已知性别用户稀疏逻辑回归配置信息，训练稀疏逻辑回归模型。

进一步的，所述应用程序中同类使用行为的统计数据可以包括下述至少一项：

消费应用程序的消费行为中对消费金额、消费频率和/或消费品分类的统计数量或比例；

阅读器应用程序的阅读行为中对阅读内容分类的统计数量或比例；

健身应用程序的健身训练行为中，对健身训练内容、健身训练强度和/或健身训练频率的统计数量或比例。

进一步的，所述性别识别模块，可以包括：

迭代决策树模型识别子模块，用于根据所述未知性别用户的第一特征数据，基于迭代决策树学习算法和算法参数，生成未知性别用户迭代决策树配置信息，应用所述迭代决策树模型，识别未知性别用户性别；

稀疏逻辑回归模型识别子模块，用于根据所述未知性别用户的第一特征数据以及第二特征数据，基于稀疏逻辑回归学习算法和算法参数，生成未知性别用户稀疏逻辑回归配置信息，应用稀疏逻辑回归模型，识别

未知性别用户性别。

进一步的，所述性别识别模块，具体可以用于：根据所述未知性别用户的特征数据，基于性别分类模型识别所述未知性别用户性别为男或女，以及对应的概率。

本发明实施例三提供的用户性别识别装置，与本发明任意实施例所提供的用户性别识别方法属于同一发明构思，可执行本发明任意实施例所提供的用户性别识别方法，具备执行用户性别识别方法相应的功能模块和有益效果。未在本实施例中详尽描述的技术细节，可参见本发明任意实施例提供的用户性别识别方法。

图 4 示出了适于用来实现本申请实施例的终端设备或服务器的计算机系统的结构示意图。

如图 4 所示，计算机系统 400 包括中央处理单元（CPU）401，其可以根据存储在只读存储器（ROM）402 中的程序或者从存储部分 408 加载到随机访问存储器（RAM）403 中的程序而执行各种适当的动作和处理。在 RAM 403 中，还存储有系统 400 操作所需的各种程序和数据。CPU401、ROM 402 以及 RAM403 通过总线 404 彼此相连。输入/输出（I/O）接口 405 也连接至总线 404。

以下部件连接至 I/O 接口 405：包括键盘、鼠标等的输入部分 406；包括诸如阴极射线管（CRT）、液晶显示器（LCD）等以及扬声器等的输出部分 407；包括硬盘等的存储部分 408；以及包括诸如 LAN 卡、调制解调器等的网络接口卡的通信部分 409。通信部分 409 经由诸如因特网的网络执行通信处理。驱动器 410 也根据需要连接至 I/O 接口 405。可拆卸介质 411，诸如磁盘、光盘、磁光盘、半导体存储器等等，根据需要安装在驱动器 410 上，以便于从其上读出的计算机程序根据需要被安装入存储部分 408。

特别地，根据本公开的实施例，上文参考流程图描述的过程可以被实现为计算机软件程序。例如，本公开的实施例包括一种计算机程序产品，其包括有形地包含在机器可读介质上的计算机程序，所述计算机程序包含用于执行流程图所示的方法的程序代码。在这样的实施例中，该计算机程序可以通过通信部分 409 从网络上被下载和安装，和/或从可拆卸介质 411 被安装。

附图中的流程图和框图，图示了按照本申请各种实施例的系统、方法和计算机程序产品的可能实现的体系架构、功能和操作。在这点上，流程图或框图中的每个方框可以代表一个模块、程序段、或代码的一部分，所述模块、程序段、或代码的一部分包含一个或多个用于实现规定的逻辑功能的可执行指令。也应当注意，在有些作为替换的实现中，方框中所标注的功能也可以以不同于附图中所标注的顺序发生。例如，两个接连地表示的方框实际上可以基本并行地执行，它们有时也可以按相反的顺序执行，这依所涉及的功能而定。也要注意的是，框图和/或流程图中的每个方框、以及框图和/或流程图中的方框的组合，可以用执行规定的功能或操作的专用的基于硬件的系统来实现，或者可以用专用硬件与计算机指令的组合来实现。

作为另一方面，本申请还提供了一种非易失性计算机存储介质，该非易失性计算机存储介质可以是上述实施例中所述装置中所包含的非易失性计算机存储介质；也可以是单独存在，未装配入终端中的非易失性计算机存储介质。上述非易失性计算机存储介质存储有一个或者多个程序，当所述一个或者多个程序被一个设备执行时，使得所述设备：获取未知性别用户的应用程序使用数据；从所述未知性别用户的应用程序使用数据中提取具有性别识别性的特征数据；根据所述未知性别用户的特征数据，基于性别分类模型对未知性别用户进行性别识别，其中，所述性别分类模型为根据已知性别用户的应用程序使用数据的特征数据进行

训练而确定。

注意，上述仅为本发明的较佳实施例及所运用技术原理。本领域技术人员会理解，本发明不限于这里所述的特定实施例，对本领域技术人员来说能够进行各种明显的变化、重新调整和替代而不会脱离本发明的保护范围。因此，虽然通过以上实施例对本发明进行了较为详细的说明，但是本发明不仅仅限于以上实施例，在不脱离本发明构思的情况下，还可以包括更多其他等效实施例，而本发明的范围由所附的权利要求范围决定。

# 权 利 要 求 书

1、一种用户性别识别方法，其特征在于，包括：

获取未知性别用户的应用程序使用数据；

从所述未知性别用户的应用程序使用数据中提取具有性别识别性的特征数据；

根据所述未知性别用户的特征数据，基于性别分类模型对未知性别用户进行性别识别，其中，所述性别分类模型为根据已知性别用户的应用程序使用数据的特征数据进行训练而确定。

2、根据权利要求 1 所述的方法，其特征在于，所述应用程序使用数据包括：

应用程序安装列表和应用程序使用行为中的至少一个。

3、根据权利要求 2 所述的方法，其特征在于，还包括：

获取已知性别用户的应用程序使用数据；

从所述已知性别用户的应用程序使用数据中提取具有性别识别性的特征数据；

根据所述特征数据，训练性别分类模型。

4、根据权利要求 3 所述的方法，其特征在于，从所述已知性别用户的应用程序使用数据中提取具有性别识别性的特征数据，包括下述至少一项:

根据所述应用程序安装列表，提取具有性别识别性的第一特征数据，

所述第一特征数据为至少两个设定类别的应用程序的数量；

根据所述应用程序安装列表，提取具有性别识别性的第二特征数据，所述第二特征数据为至少两个应用程序安装人数之间的性别比例；

根据所述应用程序使用行为，提取具有性别识别性的第三特征数据，所述第三特征数据为应用程序中同类使用行为的统计数据。

5、根据权利要求 4 所述的方法，其特征在于，根据所述特征数据，训练性别分类模型，包括：

根据所述第一特征数据，基于迭代决策树学习算法和算法参数，生成已知性别用户迭代决策树配置信息，训练迭代决策树模型；

根据所述第一特征数据以及第二特征数据，基于稀疏逻辑回归学习算法和算法参数，生成已知性别用户稀疏逻辑回归配置信息，训练稀疏逻辑回归模型。

6、根据权利要求 4 所述的方法，其特征在于，所述应用程序中同类使用行为的统计数据包括下述至少一项：

消费应用程序的消费行为中对消费金额、消费频率和/或消费品分类的统计数量或比例；

阅读器应用程序的阅读行为中对阅读内容分类的统计数量或比例；

健身应用程序的健身训练行为中，对健身训练内容、健身训练强度和/或健身训练频率的统计数量或比例。

7、根据权利要求 5 所述的方法，其特征在于，根据所述未知性别用

户的特征数据，基于性别分类模型对未知性别用户进行性别识别，包括：

根据所述未知性别用户的第一特征数据，基于迭代决策树学习算法和算法参数，生成未知性别用户迭代决策树配置信息，应用所述迭代决策树模型，识别未知性别用户性别；或

根据所述未知性别用户的第一特征数据以及第二特征数据，基于稀疏逻辑回归学习算法和算法参数，生成未知性别用户稀疏逻辑回归配置信息，应用稀疏逻辑回归模型，识别未知性别用户性别。

8、根据权利要求 1 所述的方法，其特征在于，根据所述未知性别用户的特征数据，基于性别分类模型对未知性别用户进行性别识别，包括：

根据所述未知性别用户的特征数据，基于性别分类模型识别所述未知性别用户性别为男或女，以及对应的概率。

9、一种用户性别识别装置，其特征在于，包括：

第一使用数据获取模块，用于获取未知性别用户的应用程序使用数据；

第一特征数据提取模块，用于从所述未知性别用户的应用程序使用数据中提取具有性别识别性的特征数据；

性别识别模块，用于根据所述未知性别用户的特征数据，基于性别分类模型对未知性别用户进行性别识别，其中，所述性别分类模型为根据已知性别用户的应用程序使用数据的特征数据进行训练而确定。

10、根据权利要求 9 所述的装置，其特征在于，所述应用程序使用

数据包括：

应用程序安装列表和应用程序使用行为中的至少一个。

11、根据权利要求 10 所述的装置，其特征在于，还包括：

第二使用数据获取模块，用于获取已知性别用户的应用程序使用数据；

第二特征数据提取模块，用于从所述已知性别用户的应用程序使用数据中提取具有性别识别性的特征数据；

模型训练模块，用于根据所述特征数据，训练性别分类模型。

12、根据权利要求 11 所述的装置，其特征在于，第二特征数据提取模块具体用于执行下述至少一项:

根据所述应用程序安装列表，提取具有性别识别性的第一特征数据，所述第一特征数据为至少两个设定类别的应用程序的数量；

根据所述应用程序安装列表，提取具有性别识别性的第二特征数据，所述第二特征数据为至少两个应用程序安装人数之间的性别比例；

根据所述应用程序使用行为，提取具有性别识别性的第三特征数据，所述第三特征数据为应用程序中同类使用行为的统计数据。

13、根据权利要求 12 所述的装置，其特征在于，所述模型训练模块，包括：

迭代决策树模型训练子模块，用于根据所述第一特征数据，基于迭代决策树学习算法和算法参数，生成已知性别用户迭代决策树配置信息，

训练迭代决策树模型；

稀疏逻辑回归模型训练子模块，用于根据所述第一特征数据以及第二特征数据，基于稀疏逻辑回归学习算法和算法参数，生成已知性别用户稀疏逻辑回归配置信息，训练稀疏逻辑回归模型。

14、根据权利要求 12 所述的装置，其特征在于，所述应用程序中同类使用行为的统计数据包括下述至少一项：

消费应用程序的消费行为中对消费金额、消费频率和/或消费品分类的统计数量或比例；

阅读器应用程序的阅读行为中对阅读内容分类的统计数量或比例；

健身应用程序的健身训练行为中，对健身训练内容、健身训练强度和/或健身训练频率的统计数量或比例。

15、根据权利要求 13 所述的装置，其特征在于，所述性别识别模块，包括：

迭代决策树模型识别子模块，用于根据所述未知性别用户的第一特征数据，基于迭代决策树学习算法和算法参数，生成未知性别用户迭代决策树配置信息，应用所述迭代决策树模型，识别未知性别用户性别；

稀疏逻辑回归模型识别子模块，用于根据所述未知性别用户的第一特征数据以及第二特征数据，基于稀疏逻辑回归学习算法和算法参数，生成未知性别用户稀疏逻辑回归配置信息，应用稀疏逻辑回归模型，识别未知性别用户性别。

16、根据权利要求 9 所述的装置，其特征在于，所述性别识别模块具体用于：根据所述未知性别用户的特征数据，基于性别分类模型识别所述未知性别用户性别为男或女，以及对应的概率。

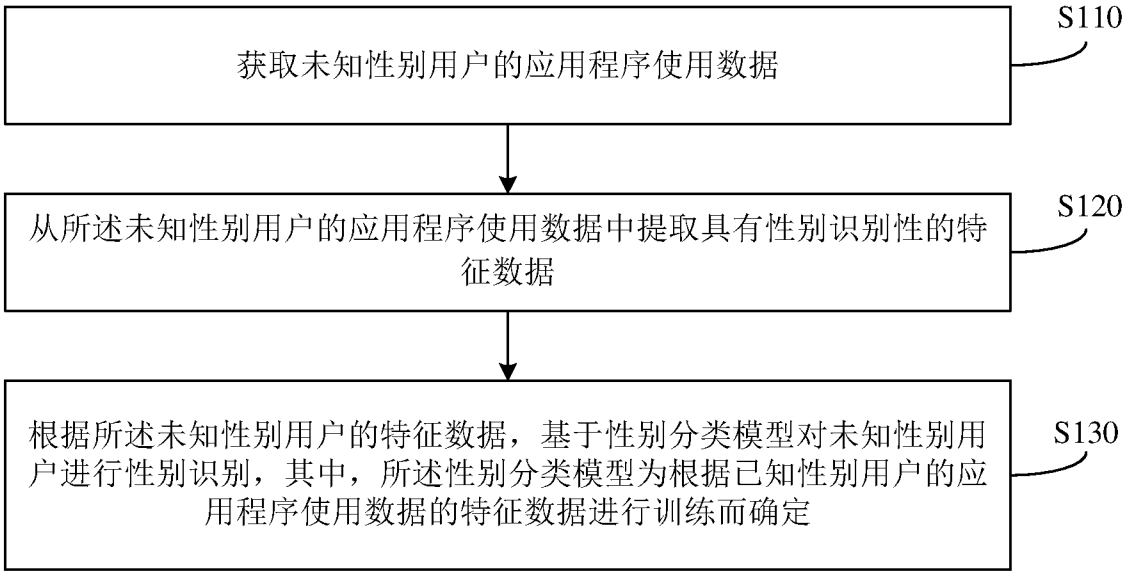

S110
获取未知性别用户的应用程序使用数据
S120
从所述未知性别用户的应用程序使用数据中提取具有性别识别性的特征数据
S130
根据所述未知性别用户的特征数据，基于性别分类模型对未知性别用户进行性别识别，其中，所述性别分类模型为根据已知性别用户的应用程序使用数据的特征数据进行训练而确定

图 1

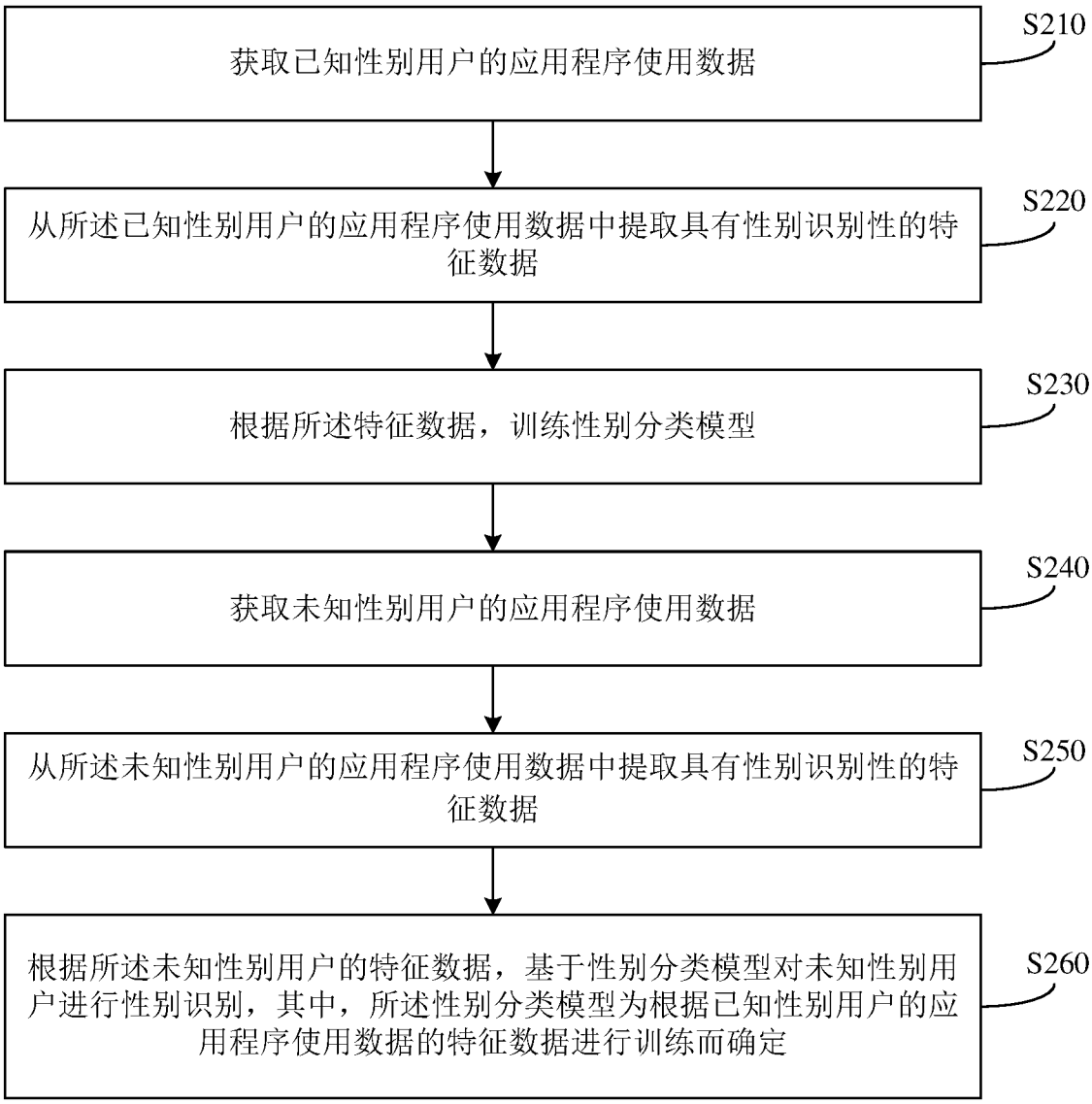

获取已知性别用户的应用程序使用数据
S210
从所述已知性别用户的应用程序使用数据中提取具有性别识别性的特征数据
S220
根据所述特征数据，训练性别分类模型
S230
获取未知性别用户的应用程序使用数据
S240
从所述未知性别用户的应用程序使用数据中提取具有性别识别性的特征数据
S250
根据所述未知性别用户的特征数据，基于性别分类模型对未知性别用户进行性别识别，其中，所述性别分类模型为根据已知性别用户的应用程序使用数据的特征数据进行训练而确定
S260

图 2

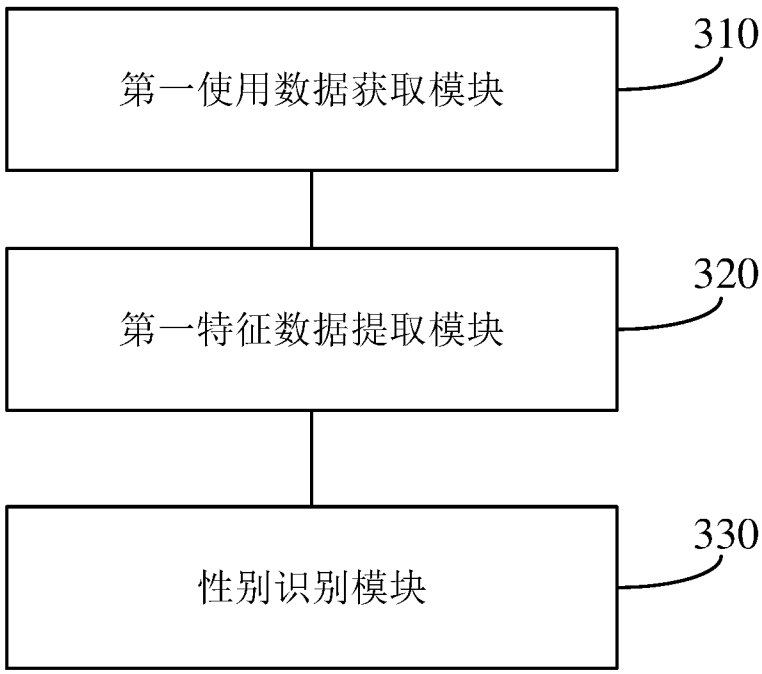

第一使用数据获取模块
310
第一特征数据提取模块
320
性别识别模块
330

图 3

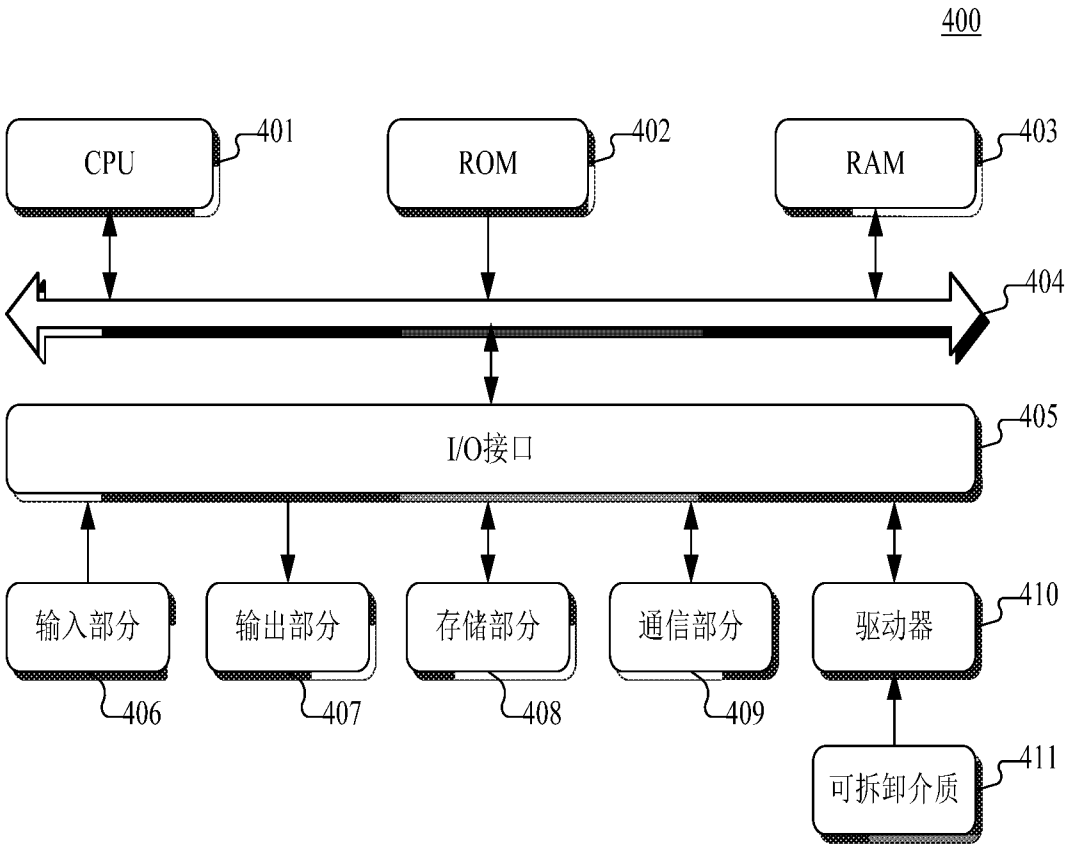

400
CPU
401
ROM
402
RAM
403
404
I/O接口
405
输入部分
406
输出部分
407
存储部分
408
通信部分
409
驱动器
410
可拆卸介质
411

图 4

# INTERNATIONAL SEARCH REPORT

International application No.
**PCT/CN2016/087685**

**A. CLASSIFICATION OF SUBJECT MATTER**

G06K 9/62 (2006.01) i

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

**B. FIELDS SEARCHED**

Minimum documentation searched (classification system followed by classification symbols)

G06K

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)

WPI; EPODOC; CNKI; CNPAT: discrimination, judgement, software, program, information, sex+, gender, identif+, recogni+, classification, application, recommend+, data, mining, mine+, analy+, process+, model+, privacy

**C. DOCUMENTS CONSIDERED TO BE RELEVANT**

| Category* | Citation of document, with indication, where appropriate, of the relevant passages | Relevant to claim No. |
|---|---|---|
| X | CN 104102819 A (BEIJING QIYI CENTURY SCIENCE AND TECHNOLOGY CO., LTD.), 15 October 2014 (15.10.2014), description, paragraphs 57-161, and figures 1-6 | 1-16 |
| X | CN 104933075 A (BAIDU ONLINE NETWORK TECHNOLOGY (BEIJING) CO., LTD.), 23 September 2015 (23.09.2015), description, paragraphs 20-74, and figures 1-2 | 1-16 |
| A | CN 103164470 A (SHANGHAI SHANDA NETWORKING CO., LTD.), 19 June 2013 (19.06.2013), the whole document | 1-16 |
| A | CN 103838884 A (LENOVO (BEIJING) CO., LTD.), 04 June 2014 (04.06.2014), the whole document | 1-16 |
| A | US 2006173556 A1 (OUTLAND RESEARCH, LLC), 03 August 2006 (03.08.2006), the whole document | 1-16 |

☐ 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 |
|---|---|
| 02 September 2016 (02.09.2016) | **28 September 2016 (28.09.2016)** |
| Name and mailing address of the ISA/CN:<br>State Intellectual Property Office of the P. R. China<br>No. 6, Xitucheng Road, Jimenqiao<br>Haidian District, Beijing 100088, China<br>Facsimile No.: (86-10) 62019451 | Authorized officer<br>**WU, Lili**<br>Telephone No.: (86-10) **62413208** |

Form PCT/ISA/210 (second sheet) (July 2009)

# INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
**PCT/CN2016/087685**

| Patent Documents referred in the Report | Publication Date | Patent Family | Publication Date |
|---|---|---|---|
| CN 104102819 A | 15 October 2014 | None | |
| CN 104933075 A | 23 September 2015 | None | |
| CN 103164470 A | 19 June 2013 | None | |
| CN 103838884 A | 04 June 2014 | None | |
| US 2006173556 A1 | 03 August 2006 | US 2006173828 A1 | 03 August 2006 |
| | | US 2007061314 A1 | 15 March 2007 |
| | | US 2007106663 A1 | 10 May 2007 |
| | | US 2007276870 A1 | 29 November 2007 |
| | | WO 2006083861 A2 | 10 August 2006 |
| | | WO 2006083861 A3 | 14 August 2008 |

# 国际检索报告

国际申请号 PCT/CN2016/087685

A. 主题的分类

G06K 9/62(2006.01)i

按照国际专利分类(IPC)或者同时按照国家分类和IPC两种分类

B. 检索领域

检索的最低限度文献(标明分类系统和分类号)

G06K

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））

WPI;EPODOC;CNKI;CNPAT: 性别、分类、识别、判别、判断、应用、软件、程序、推荐、数据、挖掘、分析、处理、信息、模型、隐私、sex+、gender、identif+、recogni+、classification、application、recommend+、data、mining、mine+、analy+、process+、model+、privacy

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 104102819 A (北京奇艺世纪科技有限公司) 2014年 10月 15日 (2014-10-15)<br>说明书第57-161段，附图1-6 | 1-16 |
| X | CN 104933075 A (百度在线网络技术北京有限公司) 2015年 9月 23日 (2015-09-23)<br>说明书第20-74段，附图1-2 | 1-16 |
| A | CN 103164470 A (盛大计算机上海有限公司) 2013年 6月 19日 (2013-06-19)<br>全文 | 1-16 |
| A | CN 103838884 A (联想北京有限公司) 2014年 6月 4日 (2014-06-04)<br>全文 | 1-16 |
| A | US 2006173556 A1 (OUTLAND RESEARCH,. LLC.) 2006年 8月 3日 (2006-08-03)<br>全文 | 1-16 |

☐ 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 |
|---|---|
| 2016年 9月 2日 | 2016年 9月 28日 |
| ISA/CN的名称和邮寄地址<br>中华人民共和国国家知识产权局(ISA/CN)<br>中国北京市海淀区蓟门桥西土城路6号 100088<br>传真号 (86-10)62019451 | 受权官员<br>吴丽丽<br>电话号码 (86-10)62413208 |

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2016/087685

| 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 | 104102819 | A | 2014年 10月 15日 | 无 | | | |
| CN | 104933075 | A | 2015年 9月 23日 | 无 | | | |
| CN | 103164470 | A | 2013年 6月 19日 | 无 | | | |
| CN | 103838884 | A | 2014年 6月 4日 | 无 | | | |
| US | 2006173556 | A1 | 2006年 8月 3日 | US | 2006173828 | A1 | 2006年 8月 3日 |
| | | | | US | 2007061314 | A1 | 2007年 3月 15日 |
| | | | | US | 2007106663 | A1 | 2007年 5月 10日 |
| | | | | US | 2007276870 | A1 | 2007年 11月 29日 |
| | | | | WO | 2006083861 | A2 | 2006年 8月 10日 |
| | | | | WO | 2006083861 | A3 | 2008年 8月 14日 |