

(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
2018 年 11 月 1 日 (01.11.2018)



(10) 国际公布号
WO 2018/196553 A1

(51) 国际专利分类号:
G06F 17/00 (2006.01)

(21) 国际申请号: PCT/CN2018/081337

(22) 国际申请日: 2018 年 3 月 30 日 (30.03.2018)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
201710290180.5 2017年4月27日 (27.04.2017) 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: 袁小燕 (YUAN, Xiaoyan); 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: 北京康信知识产权代理有限公司 (KANGXIN PARTNERS, P.C.); 中国北京市海淀区知春路甲48号盈都大厦A座16层, 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, 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,

(54) Title: METHOD AND APPARATUS FOR OBTAINING IDENTIFIER, STORAGE MEDIUM, AND ELECTRONIC DEVICE

(54) 发明名称: 标识的获取方法及装置、存储介质以及电子装置

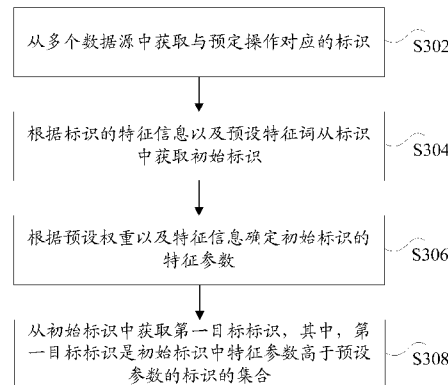


图 3

- S302 OBTAIN IDENTIFIERS CORRESPONDING TO A PREDETERMINED OPERATION FROM MULTIPLE DATA SOURCES
S304 OBTAIN INITIAL IDENTIFIERS FROM AMONG THE IDENTIFIERS ACCORDING TO CHARACTERISTIC INFORMATION OF THE IDENTIFIERS AND PRESET CHARACTERISTIC WORDS
S306 DETERMINE CHARACTERISTIC PARAMETERS OF THE INITIAL IDENTIFIERS ACCORDING TO A PRESET WEIGHT AND THE CHARACTERISTIC INFORMATION
S308 OBTAIN A FIRST TARGET IDENTIFIER FROM AMONG THE INITIAL IDENTIFIERS, WHEREIN THE FIRST TARGET IDENTIFIER IS A COLLECTION OF IDENTIFIERS OF WHICH THE CHARACTERISTIC PARAMETERS ARE HIGHER THAN A PRESET PARAMETER IN THE INITIAL IDENTIFIERS

(57) Abstract: Disclosed are a method and apparatus for obtaining an identifier, a storage medium, and an electronic device. The method comprises: obtaining identifiers corresponding to a predetermined operation from multiple data sources; obtaining initial identifiers from among the identifiers according to characteristic information of the identifiers and preset characteristic words; determining characteristic parameters of the initial identifiers according to a preset weight and the characteristic information; obtaining a first target identifier from among the initial identifiers, wherein the first target identifier is a collection of identifiers of which the characteristic parameters

PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL,
SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG,
US, UZ, VC, VN, ZA, ZM, ZW。

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

本国际公布:

— 包括国际检索报告 (条约第21条(3))。

are higher than a preset parameter in the initial identifiers. By means of the technical solution, the technical problem in the related art of low accuracy for obtaining an identifier for training is resolved.

(57) 摘要: 本发明公开了一种标识的获取方法及装置、存储介质以及电子装置。其中, 该方法包括: 从多个数据源中获取与预定操作对应的标识; 根据标识的特征信息以及预设特征词从标识中获取初始标识; 根据预设权重以及特征信息确定初始标识的特征参数; 从初始标识中获取第一目标标识, 其中, 第一目标标识是初始标识中特征参数高于预设参数的标识的集合。采用上述技术方案, 解决了相关技术中获取用于训练的标识的准确度低的技术问题。

标识的获取方法及装置、存储介质以及电子装置

本申请要求于 2017 年 04 月 27 日提交中国专利局、申请号为 201710290180.5、发明名称“标识的获取方法及装置”的中国专利申请的优先权，其全部内容通过引用结合在本申请中。

技术领域

本发明涉及计算机领域，具体而言，涉及一种标识的获取方法及装置、存储介质以及电子装置。

背景技术

在众多推荐领域，比如广告推荐、游戏推荐、视频推荐、新闻推荐等，常常需要将资源投放给某一特定领域用户（定向用户）来提升资源的投放效果，而定向用户的挖掘，通常采用训练预测模型的方式，包括逻辑回归（Logistic Regression，简称为 LR）、随机森林（Random Forest，简称为 RF）、梯度提升决策树（Gradient Boosting Decision Tree，简称为 GBDT）等，而以上任何模型效果好坏的关键，就是在其训练阶段选择训练样本（可以是用户的标识）的准确性，即正负样本选择的是否足够精准。而通常获取真实正样本的方法，是根据客户关系管理（Customer Relationship Management，简称为 CRM）获得真实可靠的正样本数据，但往往这类数据规模较小，从而导致训练出的模型特征不够明显，从而影响模型训练效果。

现有的训练数据样本获取的方式，多数是基于用户行为从单一数据源中获取规则匹配的人群，作为正样本集，负样本集则是从大盘中随机选取；这样单数据源的方式很容易导致样本有偏，同时生成的样本集规模也相对较小，除此之外选出的样本集也不易区分出每个样本的纯净度。

在现有的训练数据样本获取的方式中，如图 1 所示，根据要挖掘的特

定人群，准备样本表征词和优化规则，在单一的用户行为日志中，通过模式匹配（正则匹配）方式挖掘出带有样本表征词特征的人群，作为其训练数据正样本人群，负样本人群则是在大盘人群中排除正样本人群后，随机选择的样本。这种方式将会导致以下缺陷：首先用户行为日志单一，搜索匹配的人群有限，样本易偏；其次，正样本人群通过模式匹配挖掘后，不足以说明正样本的纯净度和可靠性。以上缺陷导致了现有的训练数据样本获取的方式获取用于训练的标识的准确度较低。

针对上述的问题，目前尚未提出有效的解决方案。

发明内容

本发明实施例提供了一种标识的获取方法及装置、存储介质以及电子装置，以至少解决相关技术中获取用于训练的标识的准确度低的技术问题。

根据本发明实施例的一个方面，提供了一种标识的获取方法，包括：从多个数据源中获取与预定操作对应的标识，其中，在所述多个数据源包括的目标数据源中记录有与所述标识对应的帐号和所述帐号执行过的所述预定操作；根据所述标识的特征信息以及预设特征词从所述标识中获取初始标识，其中，所述特征信息用于表示所述预定操作的特征；根据预设权重以及所述特征信息确定所述初始标识的特征参数，其中，所述预设权重与所述目标数据源对应，所述预设权重用于指示所述目标数据源中的帐号执行所述预定操作的频率，所述特征参数用于指示所述初始标识执行所述预定操作的频率；从所述初始标识中获取第一目标标识，其中，所述第一目标标识是所述初始标识中所述特征参数高于预设参数的标识的集合。

根据本发明实施例的另一方面，还提供了一种标识的获取装置，包括：第一获取模块，被设置为从多个数据源中获取与预定操作对应的标识，其中，在所述多个数据源包括的目标数据源中记录有与所述标识对应的帐号和所述帐号执行过的所述预定操作；第二获取模块，被设置为根据所述标识的特征信息以及预设特征词从所述标识中获取初始标识，其中，所述特

征信息用于表示所述预定操作的特征；确定模块，被设置为根据预设权重以及所述特征信息确定所述初始标识的特征参数，其中，所述预设权重与所述目标数据源对应，所述预设权重用于指示所述目标数据源中的帐号执行所述预定操作的频率，所述特征参数用于指示所述初始标识执行所述预定操作的频率；第三获取模块，被设置为从所述初始标识中获取第一目标标识，其中，所述第一目标标识是所述初始标识中所述特征参数高于预设参数的标识的集合。

根据本发明实施例的另一方面，还提供了一种存储介质，所述存储介质包括存储的程序，其中，在所述程序运行时控制所述存储介质所在设备执行上述标识的获取方法。

在本发明实施例中，从多个数据源中获取与预定操作对应的标识，其中，在多个数据源包括的目标数据源中记录有与标识对应的帐号和帐号执行过的预定操作；根据标识的特征信息以及预设特征词从标识中获取初始标识，其中，特征信息用于表示预定操作的特征；根据预设权重以及特征信息确定初始标识的特征参数，其中，预设权重与目标数据源对应，预设权重用于指示目标数据源中的帐号执行预定操作的频率，特征参数用于指示初始标识执行预定操作的频率；从初始标识中获取第一目标标识，其中，第一目标标识是初始标识中特征参数高于预设参数的标识的集合。也就是说，在目标数据源中记录了标识对应的帐号以及帐号执行过的预定操作，从中获取预定操作对应的标识，使得标识的获取途径更加的广泛，避免了从单一的用户日志获取标识规模较小导致的获取的标识有偏的问题，再根据标识的特征信息以及预设特征词初步地筛选出初始标识，并根据预设权重和特征信息为初始标识确定特征参数来表示出初始标识执行该预定操作的频率，然后从初始标识中获取特征参数高于预设参数的第一目标标志，使得第一目标标识中包括的标识均为执行预定操作频率较高的标识，从而提高了获取用于训练的标识的准确度，进而克服相关技术中获取用于训练的标识的准确度低的问题。

附图说明

此处所说明的附图用来提供对本发明的进一步理解，构成本申请的一部分，本发明的示意性实施例及其说明用于解释本发明，并不构成对本发明的不当限定。在附图中：

图 1 是根据相关技术的一种标识的获取方法的示意图；

图 2 是根据本发明实施例的一种可选的标识的获取方法的应用环境示意图；

图 3 是根据本发明实施例的一种可选的标识的获取方法的示意图；

图 4 是根据本发明实施例的一种可选的标识的获取装置的示意图一；

图 5 是根据本发明实施例的一种可选的标识的获取装置的示意图二；

图 6 是根据本发明实施例的一种可选的标识的获取装置的示意图三；

图 7 是根据本发明实施例的一种可选的标识的获取装置的示意图四；

图 8 是根据本发明实施例的一种可选的标识的获取装置的示意图五；

图 9 是根据本发明实施例的一种可选的标识的获取装置的示意图六；

图 10 是根据本发明实施例的一种可选的标识的获取方法的应用场景示意图；以及

图 11 是根据本发明实施例的一种电子装置的结构示意图。

具体实施方式

为了使本技术领域的人员更好地理解本发明方案，下面将结合本发明实施例中的附图，对本发明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发明一部分的实施例，而不是全部的实施例。基于本发明中的实施例，本领域普通技术人员在没有做出创造性劳动前提下所获得的所有其他实施例，都应当属于本发明保护的范围。

需要说明的是，本发明的说明书和权利要求书及上述附图中的术语“第一”、“第二”等是用于区别类似的对象，而不必用于描述特定的顺序或先后次序。应该理解这样使用的数据在适当情况下可以互换，以便这里描述的本发明的实施例能够以除了在这里图示或描述的那些以外的顺序实施。此外，术语“包括”和“具有”以及他们的任何变形，意图在于覆盖不排他的包含，例如，包含了一系列步骤或单元的过程、方法、系统、产品或设备不必限于清楚地列出的那些步骤或单元，而是可包括没有清楚地列出的或对于这些过程、方法、产品或设备固有的其它步骤或单元。

在本发明实施例中，提供了一种上述标识的获取方法的实施例。作为一种可选的实施方式，该标识的获取方法可以但不限于应用于如图 2 所示的应用环境中，服务器 202，被设置为从多个数据源中获取与预定操作对应的标识，根据标识的特征信息以及预设特征词从标识中获取初始标识，根据标识的特征信息以及预设特征词从标识中获取初始标识，从初始标识中获取第一目标标识；其中，在多个数据源包括的目标数据源中记录有与标识对应的帐号和帐号执行过的操作；特征信息用于表示预定操作的特征；预设权重与目标数据源对应，预设权重用于指示目标数据源中的帐号执行预定操作的频率，特征参数用于指示初始标识执行预定操作的频率；第一目标标识是初始标识中特征参数高于预设参数的标识的集合。

在本实施例中，在目标数据源中记录了标识对应的帐号以及帐号执行过的操作，服务器 202 从中获取预定操作对应的标识，使得标识的获取途径更加的广泛，避免了从单一的用户日志获取标识规模较小导致的获取的标识有偏的问题，再根据标识的特征信息以及预设特征词初步地筛选出初始标识，并根据预设权重和特征信息为初始标识确定特征参数来表示出初始标识执行该预定操作的频率，然后从初始标识中获取特征参数高于预设参数的第一目标标志，使得第一目标标识中包括的标识均为执行预定操作频率较高的标识，从而提高了获取用于训练的标识的准确度，进而克服相关技术中获取用于训练的标识的准确度低的问题。

可选地，在本实施例中，服务器 202 被设置为：获取第一特征词与第二特征词，其中，预设特征词包括第一特征词和第二特征词；从标识中获取初始标识，其中，初始标识对应的特征信息中携带第一特征词且未携带第二特征词。

可选地，在本实施例中，服务器 202 被设置为：获取预设权重，其中，预设权重的值越大表示目标数据源中的帐号执行预定操作的频率越高；从特征信息中获取时间信息和频次信息，其中，时间信息用于指示标识执行预定操作的时间，频次信息用于指示标识执行预定操作的频次；根据预设权重、时间信息以及频次信息确定特征参数，其中，特征参数的值越大表示初始标识执行预定操作的频率越高。

可选地，在本实施例中，服务器 202 被设置为：获取目标数据源中执行预定操作的帐号在目标数据源中包括的全部帐号中所占的比例；根据比例为目标数据源分配预设权重，其中，比例越大的数据源分配的预设权重越大；或者，获取第一标识集合与预设标识集合中相同标识的数量，其中，第一标识集合是初始标识中在一个目标数据源中包括的标识的集合；根据数量与第一标识集合中标识的数量之间的比值为目标数据源分配预设权重，其中，比值越大的数据源分配的预设权重越大。

可选地，在本实施例中，服务器 202 被设置为：计算初始标识在每个目标数据源中对应的时间信息和频次信息的乘积；根据预设权重计算乘积的加权和，得到特征参数。

可选地，在本实施例中，服务器 202 被设置为：从标识对应的预定操作中获取用于表示预定操作的特征的信息，其中，用于表示预定操作的特征的信息包括：预定操作对应的特征词，时间信息和频次信息；将特征词、时间信息以及频次信息存储为预设格式，得到特征信息。

可选地，在本实施例中，服务器 202 被设置为：将初始标识按照特征参数从高到低进行排列；从排列后的标识中选择出第一目标标识，其中，第一目标标识包括在排列后的标识中排在前 N 位的标识；或者，从初始标

识中获取特征参数的值大于或者等于预设值的第一目标标识。

可选地，在本实施例中，服务器 202 被设置为：将第一目标标识与预设目标标识进行匹配；在第一目标标识与预设目标标识匹配成功的情况下，确定出第一目标标识为所需的标识；在第一目标标识与预设目标标识匹配不成功的情况下，重新获取第一目标标识。

可选地，在本实施例中，服务器 202 还被设置为：判断第一目标标识与预设目标标识中是否包括大于或者等于预设数量的相同标识；在判断出第一目标标识与预设目标标识中包括大于或者等于预设数量的相同标识的情况下，确定第一目标标识与预设目标标识匹配成功。

可选地，在本实施例中，服务器 202 还被设置为：获取多个数据源中包括的帐号对应的标识；从多个数据源中包括的帐号对应的标识中随机获取除第一目标标识之外的标识，得到第二目标标识，其中，第二目标标识中包括的标识的数量与第一目标标识中包括的标识的数量相同。

可选地，在本实施例描述的应用环境中，还可以包括客户端，客户端通过网络与服务器 202 连接，服务器 202 还被设置为：根据第一目标标识和第二目标标识训练预测模型；根据预测模型从多个数据源包括的标识中为待推送资源获取待推送标识；向待推送标识对应的帐号所使用的客户端推送待推送资源。

可选地，在本实施例中，上述客户端可以包括但不限于以下至少之一：手机、平板电脑、笔记本电脑、台式 PC 机、数字电视及其他进行区域共享的硬件设备。上述网络可以包括但不限于以下至少之一：广域网、城域网、局域网。上述只是一种示例，本实施例对此不做任何限定。

根据本发明实施例，提供了一种标识的获取方法，如图 3 所示，该方法包括：

S302，从多个数据源中获取与预定操作对应的标识，其中，在多个数据源包括的目标数据源中记录有与标识对应的帐号和帐号执行过的操作；

S304, 根据标识的特征信息以及预设特征词从标识中获取初始标识, 其中, 特征信息用于表示预定操作的特征;

S306, 根据预设权重以及特征信息确定初始标识的特征参数, 其中, 预设权重与目标数据源对应, 预设权重用于指示目标数据源中的帐号执行预定操作的频率, 特征参数用于指示初始标识执行预定操作的频率;

S308, 从初始标识中获取第一目标标识, 其中, 第一目标标识是初始标识中特征参数高于预设参数的标识的集合。

可选地, 在本实施例中, 上述标识的获取方法可以但不限于应用于获取标识样本进行模型训练, 利用训练结果为客户端推送资源的场景中。其中, 上述客户端可以但不限于为各种类型的软件, 例如, 搜索软件、社交软件、即时通讯软件、新闻资讯软件、游戏软件、购物软件等。可选的, 可以但不限于应用于在上述获取标识样本进行模型训练, 利用训练结果为购物软件的客户端推送资源的场景中, 或还可以但不限于应用于在上述获取标识样本进行模型训练, 利用训练结果为搜索软件的客户端推送资源的场景中, 以实现标识样本的获取。上述仅是一种示例, 本实施例中对此不做任何限定。

可选地, 在本实施例中, 多个数据源可以是各种平台、软件、网站、应用程序等。例如: 社交应用、搜索引擎、电商网站、广告平台等。

可选地, 在本实施例中, 标识在不同数据源中可以对应不同的帐号。举例来说, 一个用户可能在多个应用上都注册了帐号, 例如: 在社交平台上注册了帐号 A, 在购物网站上注册了帐号 B, 在即时通讯应用上注册了帐号 C, 该用户可以将上述平台上的三个帐号关联起来, 那么, 上述三个帐号 A、B、C 就可以对应同一个标识用来唯一标识该用户。

可选地, 在本实施例中, 上述目标数据源中可以包括一个或者多个数据源。也就是说, 数据源中记录了标识对应的该数据源中的账号, 以及该帐号执行过的操作。与预定操作对应的标识可能记录在多个数据源中一个

数据源里，还可能记录在多个数据源中的几个数据源里。

可选地，在本实施例中，预定操作可以是标识执行过的某个行为或者用于表征该行为的词组。例如：如果要挖掘的用户是购买母婴类产品的用户，那么预定操作可以是“点击带有奶粉或者纸尿裤的词条”，或者“奶粉”、“纸尿裤”等词组。从多个数据源中获取的与预定操作对应的标识可以首先获取搜索引擎中搜索过“奶粉”、“纸尿裤”的帐号，购物网站中购买过奶粉或者纸尿裤的帐号，即时通讯软件中发送过带有“奶粉”、“纸尿裤”等词组的消息的帐号以及在多个数据源中点击过带有奶粉或者纸尿裤的词条的帐号，再获取上述这些帐号对应的标识。

可选地，在本实施例中，初始标识中可以但不限于包括一个或者多个标识。预设特征词可以但不限于是一个或者多个特征词。第一目标标识中可以但不限于包括一个或者多个标识。

可选地，在本实施例中，预设权重可以用于指示目标数据源中的帐号执行预定操作的频率。换句话说，预设权重可以用来表示目标数据源中的帐号对预定操作的关注程度，这个关注程度可以但不限于用目标数据源中的帐号执行预定操作的频率来表示。在这里，目标数据源中的帐号执行预定操作的频率可以但不限于指目标数据源中的帐号有多少是经常执行该预定操作的（比如：频率超过每天5次执行该预定操作的帐号占目标数据源中总帐号数的50%）。或者还可以但不限于用目标数据源中的帐号执行预定操作的显著性来表示目标数据源中的帐号执行预定操作的频率。目标数据源中的帐号执行预定操作的显著性可以通过计算初始标识中在目标数据源中记录有帐号的标识在历史数据（比如：上一次推送资源的标识）中所占的比例来确定。

可选地，在本实施例中，预设权重可以是根据目标数据源中的帐号执行预定操作的频率为目标数据源设置的，还可以是根据目标数据源中的帐号执行预定操作的频率通过模型训练的方式计算得到的。

可见，通过上述步骤，在目标数据源中记录了标识对应的帐号以及帐

号执行过的操作，从中获取预定操作对应的标识，使得标识的获取途径更加的广泛，避免了从单一的用户日志获取标识规模较小导致的获取的标识有偏的问题，再根据标识的特征信息以及预设特征词初步地筛选出初始标识，并根据预设权重和特征信息为初始标识确定特征参数来表示出初始标识执行该预定操作的频率，然后从初始标识中获取特征参数高于预设参数的第一目标标志，使得第一目标标识中包括的标识均为执行预定操作频率较高的标识，从而提高了获取用于训练的标识的准确度，进而克服相关技术中获取用于训练的标识的准确度低的问题。

作为一种可选的方案，根据标识的特征信息以及预设特征词从标识中获取初始标识包括：

S1，获取第一特征词与第二特征词，其中，预设特征词包括第一特征词和第二特征词；

S2，从标识中获取初始标识，其中，初始标识对应的特征信息中携带第一特征词且未携带第二特征词。

可选地，在本实施例中，预设特征词可以但不限于包括第一特征词和第二特征词。预设特征词可以用来表示一类用户人群的特征，其可以包括正向表征词和负向表征词，其中，正向表征词（相当于上述第一特征词），即通俗意义上的关键词（keywords），用来表征特征人群，负向表征词（相当于上述第二特征词），即过滤词（filter_words），负向表征词的作用，在于去噪，即去掉某些多词拼接后的噪声，从而让正向表征词更能表征特征人群。

通过上述步骤，根据标识的特征信息以及预设特征词中包括的第一特征词和第二特征词从标识中获取初始标识，实现了对标识的初步筛选。

作为一种可选的方案，根据预设权重以及特征信息确定初始标识的特征参数包括：

S1，获取预设权重，其中，预设权重的值越大表示目标数据源中的帐

号执行预定操作的频率越高;

S2, 从特征信息中获取时间信息和频次信息, 其中, 时间信息用于指示标识执行预定操作的时间, 频次信息用于指示标识执行预定操作的频次;

S3, 根据预设权重、时间信息以及频次信息确定特征参数, 其中, 特征参数的值越大表示初始标识执行预定操作的频率越高。

可选地, 在本实施例中, 可以但不限于通过以下方式之一获取预设权重:

方式一, 获取目标数据源中执行预定操作的帐号在目标数据源中包括的全部帐号中所占的比例; 根据比例为目标数据源分配预设权重, 其中, 比例越大的数据源分配的预设权重越大。

例如, 目标数据源有三个, 分别是目标数据源 A、目标数据源 B 和目标数据源 C, 在目标数据源 A 中共有 100 个帐号, 其中有 34 个帐号执行过预定操作, 在目标数据源 B 中共有 200 个帐号, 其中有 25 个帐号执行过预定操作, 在目标数据源 C 中共有 100 个帐号, 其中有 56 个帐号执行过预定操作。那么, 获取到目标数据源 A、目标数据源 B 和目标数据源 C 对应的比例分别为 34%、12.5%和 56%, 根据获取到的比为目标数据源 A、目标数据源 B 和目标数据源 C 分别分配的预设权重 2、1、3。

方式二, 获取第一标识集合与预设标识集合中相同标识的数量, 其中, 第一标识集合是初始标识中在一个目标数据源中包括的标识的集合; 根据数量与第一标识集合中标识的数量之间的比值为目标数据源分配预设权重, 其中, 比值越大的数据源分配的预设权重越大。

可选地, 在本实施例中, 预设标识集合可以但不限于指前一次获取的第一目标标识中目标数据源包括的标识, 或者是根据前一次推送数据的标识中目标数据源包括的标识。

在一个可选的实施方式中, 预设标识集合以前一次获取的第一目标标识中目标数据源包括的标识为例, 目标数据源 A 对应的预设标识集合 A

中包括 40 个标识, 目标数据源 B 对应的预设标识集合 B 中包括 30 个标识, 目标数据源 C 对应的预设标识集合 C 中包括 40 个标识; 初始标识中包括的来自目标数据源 A、目标数据源 B 和目标数据源 C 的标识的数量分别是 20、40、40, 那么, 目标数据源 A 对应的第一标识集合 A 中包括 20 个标识, 目标数据源 B 对应的第一标识集合 B 中包括 40 个标识, 目标数据源 C 对应的第一标识集合 C 中包括 40 个标识, 其中, 将第一标识集合 A 与预设标识集合 A 中的标识进行匹配, 获取到第一标识集合 A 与预设标识集合 A 中相同标识的数量为 10, 将第一标识集合 B 与预设标识集合 B 中的标识进行匹配, 获取到第一标识集合 B 与预设标识集合 B 中相同标识的数量为 5, 将第一标识集合 C 与预设标识集合 C 中的标识进行匹配, 获取到第一标识集合 C 与预设标识集合 C 中相同标识的数量为 20, 根据获取到的上述相同标识的数量为目标数据源 A、目标数据源 B 和目标数据源 C 分别分配的预设权重 2、1、3。

可选地, 在本实施例中, 可以通过以下方式确定特征参数: 计算初始标识在每个目标数据源中对应的时间信息和频次信息的乘积, 再根据预设权重计算乘积的加权和, 得到特征参数。

在一个可选的实施方式中, 可以通过以下公式计算上述特征参数:

$$score_i = \sum_{source=0}^n weight_{source} * exp^{-ln2*time} * sigmoid(action_{cnt})$$

其中, source 代表的是数据源, 这里有 n 个数据源; weight 代表的是每个数据源上的预设权重; time 代表的是上述时间信息, 可以用 $abs(\text{用户行为发生时间}-\text{当前挖掘时间})$, 即行为时间差的绝对值来表示上述时间信息, 其作为用户行为时间衰减参数, 即行为发生距离当前时间越近, 则其特征参数越大, 距离当前时间越远, 特征参数越小; action 代表上述频次信息, 可以用来表示用户行为频次, 这里取了 sigmoid 函数, 对其做了归一化处理; 其表示行为频次越多, 特征参数越高。

可见, 通过上述步骤, 根据预设权重以及特征信息确定初始标识的特

征参数，为初始标识打分，可以用来衡量初始标识执行预定操作的频率，从而从初始标识中筛选出的第一目标标识更能代表预定操作，从而提高了获取用于训练的标识的准确度，进而克服相关技术中获取用于训练的标识的准确度低的问题。

作为一种可选的方案，在根据标识的特征信息以及预设特征词从标识中获取初始标识之前，还包括：

S1，从标识对应的预定操作中获取用于表示预定操作的特征的信息，其中，用于表示预定操作的特征的信息包括：预定操作对应的特征词，时间信息和频次信息；

S2，将特征词、时间信息以及频次信息存储为预设格式，得到特征信息。

可见，通过上述步骤，将从标识对应的预定操作中获取的用于表示预定操作的特征的信息整理为预定格式进行存储，从而使得特征词的比对更加快速便捷。

作为一种可选的方案，从初始标识中获取第一目标标识包括以下之一：

S1，将初始标识按照特征参数从高到低进行排列；从排列后的标识中选择出第一目标标识，其中，第一目标标识包括在排列后的标识中排在前N位的标识；

S2，从初始标识中获取特征参数的值大于或者等于预设值的第一目标标识。

可选地，在本实施例中，可以对特征参数进行从高到低的排序，将排在前N位的标识作为特征参数高于预设参数的标识，得到第一目标标识。

可选地，在本实施例中，可以设定预设值，将值大于或者等于该预设值的特征参数对应的标识作为第一目标标识。

可见，通过上述步骤，通过对特征参数进行从高到低的排序，或者，

设定预设值的方式获取第一目标标识可以清楚地从初始标识中选择出更能代表预定操作的标识。

作为一种可选的方案，在从初始标识中获取第一目标标识之后，还包括：

S1，将第一目标标识与预设目标标识进行匹配；

S2，在第一目标标识与预设目标标识匹配成功的情况下，确定出第一目标标识为所需的标识；在第一目标标识与预设目标标识匹配不成功的情况下，重新获取第一目标标识。

可选地，在本实施例中，可以通过以下方式对第一目标标识与预设目标标识进行匹配：判断第一目标标识与预设目标标识中是否包括大于或者等于预设数量的相同标识，并在判断出第一目标标识与预设目标标识中包括大于或者等于预设数量的相同标识的情况下，确定第一目标标识与预设目标标识匹配成功。

可选地，在本实施例中，预设目标标识可以是上一次获取的第一目标标识，还可以是预先设定的目标标识。

可选地，在本实施例中，重新获取第一目标标识时可以但不限于通过重新设定预定操作来重新获取预定操作对应的标识从而获取第一目标标识。还可以但不限于通过重新为目标数据源分配预设权重来重新获取第一目标标识。

可见，通过上述步骤，将第一目标标识与预设目标标识进行匹配，如果匹配成功了则可以确定当前获取的第一目标标识满足模型训练的需要，也就是说，第一目标标识是所需的标识。反之，如果匹配不成功，则说明当前获取的第一目标标识不满足模型训练的需要，可以重新获取第一目标标识。

作为一种可选的方案，在从初始标识中获取第一目标标识之后，还包括：

S1, 获取多个数据源中包括的帐号对应的标识;

S2, 从多个数据源中包括的帐号对应的标识中随机获取除第一目标标识之外的标识, 得到第二目标标识, 其中, 第二目标标识中包括的标识的数量与第一目标标识中包括的标识的数量相同。

可选地, 在本实施例中, 第一目标标识可以作为模型训练的正样本, 在获取了第一目标标识之后, 还可以从多个数据源的全部的标识中获取第二目标标识作为模型训练的负样本。

作为一种可选的方案, 在从多个数据源中包括的帐号对应的标识中随机获取除第一目标标识之外的标识, 得到第二目标标识之后, 还包括:

S1, 根据第一目标标识和第二目标标识训练预测模型;

S2, 根据预测模型从多个数据源包括的标识中为待推送资源获取待推送标识;

S3, 向待推送标识推送待推送资源。

可选地, 在本实施例中, 获取的第一目标标识和第二目标标识可以用来进行预测模型的训练, 从而使得通过预测模型获取的待推送标识能够更加准确的代表预定操作所指向的人群。从而能够使得推送资源的效率能够更高。

需要说明的是, 对于前述的各方法实施例, 为了简单描述, 故将其都表述为一系列的动作组合, 但是本领域技术人员应该知悉, 本发明并不受所描述的动作顺序的限制, 因为依据本发明, 某些步骤可以采用其他顺序或者同时进行。其次, 本领域技术人员也应该知悉, 说明书中所描述的实施例均属于优选实施例, 所涉及的动作和模块并不一定是本发明所必须的。

通过以上的实施方式的描述, 本领域的技术人员可以清楚地了解到根据上述实施例的方法可借助软件加必需的通用硬件平台的方式来实现, 当然也可以通过硬件, 但很多情况下前者是更佳的实施方式。基于这样的理

解，本发明的技术方案本质上或者说对相关技术做出贡献的部分可以以软件产品的形式体现出来，该计算机软件产品存储在一个存储介质（如ROM/RAM、磁碟、光盘）中，包括若干指令用以使得一台终端设备（可以是手机，计算机，服务器，或者网络设备等）执行本发明各个实施例所述的方法。

根据本发明实施例，还提供了一种用于实施上述标识的获取方法的标识的获取装置，如图4所示，该装置包括：

1) 第一获取模块42，被设置为从多个数据源中获取与预定操作对应的标识，其中，在多个数据源包括的目标数据源中记录有与标识对应的帐号和帐号执行过的操作；

2) 第二获取模块44，被设置为根据标识的特征信息以及预设特征词从标识中获取初始标识，其中，特征信息用于表示预定操作的特征；

3) 确定模块46，被设置为根据预设权重以及特征信息确定初始标识的特征参数，其中，预设权重与目标数据源对应，预设权重用于指示目标数据源中的帐号执行预定操作的频率，特征参数用于指示初始标识执行预定操作的频率；

4) 第三获取模块48，被设置为从初始标识中获取第一目标标识，其中，第一目标标识是初始标识中特征参数高于预设参数的标识的集合。

可选地，在本实施例中，上述标识的获取装置可以但不限于应用于获取标识样本进行模型训练，利用训练结果为客户端推送资源的场景中。其中，上述客户端可以但不限于为各种类型的软件，例如，搜索软件、社交软件、即时通讯软件、新闻资讯软件、游戏软件、购物软件等。可选的，可以但不限于应用于在上述获取标识样本进行模型训练，利用训练结果为购物软件的客户端推送资源的场景中，或还可以但不限于应用于在上述获取标识样本进行模型训练，利用训练结果为搜索软件的客户端推送资源的场景中，以实现标识样本的获取。上述仅是一种示例，本实施例中对此不

做任何限定。

可选地，在本实施例中，多个数据源可以是各种平台、软件、网站、应用程序等。例如：社交应用、搜索引擎、电商网站、广告平台等。

可选地，在本实施例中，标识在不同数据源中可以对应不同的帐号。举例来说，一个用户可能在多个应用上都注册了帐号，例如：在社交平台上注册了帐号 A，在购物网站上注册了帐号 B，在即时通讯应用上注册了帐号 C，该用户可以将上述平台上的三个帐号关联起来，那么，上述三个帐号 A、B、C 就可以对应同一个标识用来唯一标识该用户。

可选地，在本实施例中，上述目标数据源中可以包括一个或者多个数据源。也就是说，数据源中记录了标识对应的该数据源中的帐号，以及该帐号执行过的操作。与预定操作对应的标识可能记录在多个数据源中一个数据源里，还可能记录在多个数据源中的几个数据源里。

可选地，在本实施例中，预定操作可以是标识执行过的某个行为或者用于表征该行为的词组。例如：如果要挖掘的用户是购买母婴类产品的用户，那么预定操作可以是“点击带有奶粉或者纸尿裤的词条”，或者“奶粉”、“纸尿裤”等词组。从多个数据源中获取的与预定操作对应的标识可以首先获取搜索引擎中搜索过“奶粉”、“纸尿裤”的帐号，购物网站中购买过奶粉或者纸尿裤的帐号，即时通讯软件中发送过带有“奶粉”、“纸尿裤”等词组的消息的帐号以及在多个数据源中点击过带有奶粉或者纸尿裤的词条的帐号，再获取上述这些帐号对应的标识。

可选地，在本实施例中，初始标识中可以但不限于包括一个或者多个标识。预设特征词可以但不限于是一个或者多个特征词。第一目标标识中可以但不限于包括一个或者多个标识。

可选地，在本实施例中，预设权重可以用于指示目标数据源中的帐号执行预定操作的频率。换句话说，预设权重可以用来表示目标数据源中的帐号对预定操作的关注程度，这个关注程度可以但不限于用目标数据源中

的帐号执行预定操作的频率来表示。在这里，目标数据源中的帐号执行预定操作的频率可以但不限于指目标数据源中的帐号有多少是经常执行该预定操作的（比如：频率超过每天5次执行该预定操作的账号占目标数据源中总账号数的50%）。或者还可以但不限于用目标数据源中的帐号执行预定操作的显著性来表示目标数据源中的帐号执行预定操作的频率。目标数据源中的帐号执行预定操作的显著性可以通过计算初始标识中在目标数据源中记录有帐号的标识在历史数据（比如：上一次推送资源的标识）中所占的比例来确定。

可选地，在本实施例中，预设权重可以是根据目标数据源中的帐号执行预定操作的频率为目标数据源设置的，还可以是根据目标数据源中的帐号执行预定操作的频率通过模型训练的方式计算得到的。

可见，通过上述装置，在目标数据源中记录了标识对应的帐号以及帐号执行过的操作，从中获取预定操作对应的标识，使得标识的获取途径更加的广泛，避免了从单一的用户日志获取标识规模较小导致的获取的标识有偏的问题，再根据标识的特征信息以及预设特征词初步地筛选出初始标识，并根据预设权重和特征信息为初始标识确定特征参数来表示出初始标识执行该预定操作的频率，然后从初始标识中获取特征参数高于预设参数的第一目标标识，使得第一目标标识中包括的标识均为执行预定操作频率较高的标识，从而提高了获取用于训练的标识的准确度，进而克服相关技术中获取用于训练的标识的准确度低的问题。

作为一种可选的方案，如图5所示，第二获取模块44包括：

1) 第一获取单元52，被设置为获取第一特征词与第二特征词，其中，预设特征词包括第一特征词和第二特征词；

2) 第二获取单元54，被设置为从标识中获取初始标识，其中，初始标识对应的特征信息中携带第一特征词且未携带第二特征词。

可选地，在本实施例中，预设特征词可以但不限于包括第一特征词和

第二特征词。预设特征词可以用来表示一类用户人群的特征，其可以包括正向表征词和负向表征词，其中，正向表征词（相当于上述第一特征词），即通俗意义上的关键词（keywords），用来表征特征人群，负向表征词（相当于上述第二特征词），即过滤词（filter_words），负向表征词的作用，在于去噪，即去掉某些多词拼接后的噪声，从而让正向表征词更能表征特征人群。

通过上述装置，根据标识的特征信息以及预设特征词中包括的第一特征词和第二特征词从标识中获取初始标识，实现了对标识的初步筛选。

作为一种可选的方案，如图6所示，确定模块46包括：

1）第三获取单元62，被设置为获取预设权重，其中，预设权重的值越大表示目标数据源中的帐号执行预定操作的频率越高；

2）第四获取单元64，被设置为从特征信息中获取时间信息和频次信息，其中，时间信息用于指示标识执行预定操作的时间，频次信息用于指示标识执行预定操作的频次；

3）确定单元66，被设置为根据预设权重、时间信息以及频次信息确定特征参数，其中，特征参数的值越大表示初始标识执行预定操作的频率越高。

可选地，在本实施例中，第三获取单元62被设置为以下之一：

获取目标数据源中执行预定操作的帐号在目标数据源中包括的全部帐号中所占的比例；根据比例为目标数据源分配预设权重，其中，比例越大的数据源分配的预设权重越大；

获取第一标识集合与预设标识集合中相同标识的数量，其中，第一标识集合是初始标识中在一个目标数据源中包括的标识的集合；根据数量与第一标识集合中标识的数量之间的比值为目标数据源分配预设权重，其中，比值越大的数据源分配的预设权重越大。

例如，目标数据源有三个，分别是目标数据源 A、目标数据源 B 和目标数据源 C，在目标数据源 A 中共有 100 个帐号，其中有 34 个帐号执行过预定操作，在目标数据源 B 中共有 200 个帐号，其中有 25 个帐号执行过预定操作，在目标数据源 C 中共有 100 个帐号，其中有 56 个帐号执行过预定操作。那么，获取到目标数据源 A、目标数据源 B 和目标数据源 C 对应的比例分别为 34%、12.5%和 56%，根据获取到的比为目标数据源 A、目标数据源 B 和目标数据源 C 分别分配的预设权重 2、1、3。

可选地，在本实施例中，预设标识集合可以但不限于指前一次获取的第一目标标识中目标数据源包括的标识，或者是根据前一次推送数据的标识中目标数据源包括的标识。

在一个可选的实施方式中，预设标识集合以前一次获取的第一目标标识中目标数据源包括的标识为例，目标数据源 A 对应的预设标识集合 A 中包括 40 个标识，目标数据源 B 对应的预设标识集合 B 中包括 30 个标识，目标数据源 C 对应的预设标识集合 C 中包括 40 个标识；初始标识中包括的来自目标数据源 A、目标数据源 B 和目标数据源 C 的标识的数量分别是 20、40、40，那么，目标数据源 A 对应的第一标识集合 A 中包括 20 个标识，目标数据源 B 对应的第一标识集合 B 中包括 40 个标识，目标数据源 C 对应的第一标识集合 C 中包括 40 个标识，其中，将第一标识集合 A 与预设标识集合 A 中的标识进行匹配，获取到第一标识集合 A 与预设标识集合 A 中相同标识的数量为 10，将第一标识集合 B 与预设标识集合 B 中的标识进行匹配，获取到第一标识集合 B 与预设标识集合 B 中相同标识的数量为 5，将第一标识集合 C 与预设标识集合 C 中的标识进行匹配，获取到第一标识集合 C 与预设标识集合 C 中相同标识的数量为 20，根据获取到的上述相同标识的数量为目标数据源 A、目标数据源 B 和目标数据源 C 分别分配的预设权重 2、1、3。

可选地，在本实施例中，第四获取单元 64 被设置为：计算初始标识在每个目标数据源中对应的时间信息和频次信息的乘积；根据预设权重计

算乘积的加权和，得到特征参数。

在一个可选的实施方式中，可以通过以下公式计算上述特征参数：

$$score_i = \sum_{source=0}^n weight_{source} * exp^{-ln2*time} * sigmoid(action_{cnt})$$

其中，source 代表的是数据源，这里有 n 个数据源；weight 代表的是每个数据源上的预设权重；time 代表的是上述时间信息，可以用 $abs(\text{用户行为发生时间}-\text{当前挖掘时间})$ ，即行为时间差的绝对值来表示上述时间信息，其作为用户行为时间衰减参数，即行为发生距离当前时间越近，则其特征参数越大，距离当前时间越远，特征参数越小；action 代表上述频次信息，可以用来表示用户行为频次，这里取了 sigmoid 函数，对其做了归一化处理；其表示行为频次越多，特征参数越高。

可见，通过上述装置，根据预设权重以及特征信息确定初始标识的特征参数，为初始标识打分，可以用来衡量初始标识执行预定操作的频率，从而从初始标识中筛选出的第一目标标识更能代表预定操作，从而提高了获取用于训练的标识的准确度，进而克服相关技术中获取用于训练的标识的准确度低的问题。

可选地，在本实施例中，该装置还包括：

第六获取模块，被设置为从标识对应的预定操作中获取用于表示预定操作的特征的信息，其中，用于表示预定操作的特征的信息包括：预定操作对应的特征词，时间信息和频次信息；

存储模块，被设置为将特征词、时间信息以及频次信息存储为预设格式，得到特征信息。

可见，通过上述装置，将从标识对应的预定操作中获取的用于表示预定操作的特征的信息整理为预定格式进行存储，从而使得特征词的比对更加快速便捷。

作为一种可选的方案,如图 7 所示,第三获取模块 48 包括以下之一:

1) 处理单元 72, 被设置为将初始标识按照特征参数从高到低进行排列;从排列后的标识中选择出第一目标标识,其中,第一目标标识包括在排列后的标识中排在前 N 位的标识;

2) 第五获取单元 74, 被设置为从初始标识中获取特征参数的值大于或者等于预设值的第一目标标识。

可选地,在本实施例中,可以对特征参数进行从高到低的排序,将排在前 N 位的标识作为特征参数高于预设参数的标识,得到第一目标标识。

可选地,在本实施例中,可以设定预设值,将值大于或者等于该预设值的特征参数对应的标识作为第一目标标识。

可见,通过上述装置,通过对特征参数进行从高到低的排序,或者,设定预设值的方式获取第一目标标识可以清楚地从初始标识中选择出更能代表预定操作的标识。

作为一种可选的方案,如图 8 所示,上述装置还包括:

1) 匹配模块 82, 被设置为将第一目标标识与预设目标标识进行匹配;

2) 处理模块 84, 被设置为在第一目标标识与预设目标标识匹配成功的情况下,确定出第一目标标识为所需的标识;在第一目标标识与预设目标标识匹配不成功的情况下,重新获取第一目标标识。

可选地,在本实施例中,匹配模块 82 被设置为:判断第一目标标识与预设目标标识中是否包括大于或者等于预设数量的相同标识;在判断出第一目标标识与预设目标标识中包括大于或者等于预设数量的相同标识的情况下,确定第一目标标识与预设目标标识匹配成功。

可选地,在本实施例中,预设目标标识可以是上一次获取的第一目标标识,还可以是预先设定的目标标识。

可选地,在本实施例中,重新获取第一目标标识时可以但不限于通过

重新设定预定操作来重新获取预定操作对应的标识从而获取第一目标标识。还可以但不限于通过重新为目标数据源分配预设权重来重新获取第一目标标识。

可见，通过上述装置，将第一目标标识与预设目标标识进行匹配，如果匹配成功了则可以确定当前获取的第一目标标识满足模型训练的需要，也就是说，第一目标标识是所需的标识。反之，如果匹配不成功，则说明当前获取的第一目标标识不满足模型训练的需要，可以重新获取第一目标标识。

作为一种可选的方案，如图9所示，上述装置还包括：

1) 第四获取模块92，被设置为获取多个数据源中包括的帐号对应的标识；

2) 第五获取模块94，被设置为从多个数据源中包括的帐号对应的标识中随机获取除第一目标标识之外的标识，得到第二目标标识，其中，第二目标标识中包括的标识的数量与第一目标标识中包括的标识的数量相同。

可选地，在本实施例中，第一目标标识可以作为模型训练的正样本，在获取了第一目标标识之后，还可以从多个数据源的全部的标识中获取第二目标标识作为模型训练的负样本。

可选地，在本实施例中，上述装置还包括：

训练模块，被设置为根据第一目标标识和第二目标标识训练预测模型；

第七获取模块，被设置为根据预测模型从多个数据源包括的标识中为待推送资源获取待推送标识；

推送模块，被设置为向待推送标识推送待推送资源。

可选地，在本实施例中，获取的第一目标标识和第二目标标识可以用来进行预测模型的训练，从而使得通过预测模型获取的待推送标识能够更

加准确的代表预定操作所指向的人群。从而能够使得推送资源的效率能够更高。

本发明实施例的应用环境可以但不限于参照实施例 1 中的应用环境，本实施例中对此不再赘述。本发明实施例提供了用于实施上述标识的获取方法的一种可选的具体应用示例。

作为一种可选的实施例，上述标识的获取方法可以但不限于应用于如图 10 所示的对标识进行获取的场景中。多个数据源为服务器提供数据，服务器根据从数据源得到的数据进行第一目标标识和第二目标标识的获取，再根据第一目标标识和第二目标标识进行预测模型的训练，通过训练好的预测模型从全部标识中筛选出待推送资源的标识，将待推送资源推送给筛选出的标识登录的客户端。

在一个可选的实施方式中，多个数据源可以包括社交/搜索/电商/广告/移动应用程序（application，简称为 app）等领域，以使用标识的用户在社交/搜索/电商/广告/移动 app 等领域的用户行为作为标识的特征信息，通过文本语义挖掘各个垂直行业上的初选人群；通过第一标识集合与预设标识集合中相同标识的匹配验证目标数据源中历史效果的显著性得到预设权重，并根据预设权重以及频率信息（例如：用户行为频度）和时间信息（例如：时间衰减因子），为初选标识排序；通过选定排在前 N 位的标识得到第一目标标识，通过第一目标标识与预设目标标识的匹配进行历史效果显著性的交叉验证，可有效选定训练数据的正样本；在大盘活跃人群中减去上述选定的正样本集合，从剩余集合中随机获取相同规模大小的第二目标标识作为负样本集合。从而实现服务器对第一目标标识和第二目标标识的获取。

在本实施方式中，通过文本语义特征挖掘获取训练数据正负样本，融合了用户在社交/搜索/电商/广告/移动 app 等领域的多种用户行为特征，然后通过用户行为频次因子（即上述频次信息）和行为时间衰减因子（即上

述时间信息), 以及用户在不同行为上的历史效果验证, 给予用户不同的行为权重因子 (即上述预设权重), 综合以上各要素, 给用户做打分 (即上述获取的特征参数) 并排序, 进而可以根据分值排序, 有效判定正样本 (即上述第一目标标识) 的纯净度, 并根据需要自由选择排位在前 N 位的标识作为训练数据正样本。从而解决了用户行为单一, 以及正样本纯净度低的问题。

在本实施例中, 能够融合用户在互联网多种场景的行为特征, 挖掘出具体特定表征意义的用户人群对应的标识, 并通过校验检测, 获得纯净度较高的正负样本。

为实现上述要求, 本实施例中的上述服务器可以包括以下功能模块:

1) 特征表征词收集模块, 被设置为根据需要筛选的特定人群对应的标识的特征定义其特征表征词 (相当于上述预设特征词), 其包括正向表征词 (相当于上述第一特征词) 和负向表征词 (相当于上述第二特征词), 其中正向表征词, 即通俗意义上的关键词 (keywords), 负向表征词, 即过滤词 (filter_words), 负向表征词的作用, 在于去噪, 即去掉某些多词拼接后的噪声, 从而让正向表征词更能表征我们的特征人群。

2) 用户多种行为特征融合模块, 被设置为通过用户在社交/搜索/电商/广告/移动 app 等领域的多种行为表述, 从中提炼 (用户标识-特征表述串-时间信息-频次信息) 这几个关键元素。

3) 模式匹配模块, 被设置为根据特征表征词收集模块中的特征表征词, 在用户多种行为特征融合模块中的用户多种行为数据 (用户标识-特征表述串-时间信息-频次信息) 中, 通过模式匹配方式, 去搜索含有正向表征词, 但不含有负向表征词的用户标识作为初选标识。

4) 用户打分模块, 被设置为对模式匹配模块中的初选标识进行打分 (即获取特征参数), 打分涉及两部分, 一部分是对数据源的预设权重 (weight) 进行计算, 一部分是细化到每个数据源内部, 计算每个初选标

识的行为分值；其中 $weight$ 的计算，有两种方式，一是分数据源切分人群包，通过第一标识集合与预设标识集合中相同标识的匹配分别验证单个目标数据源上人群包的显著性，根据显著性的相对值，来分配当前数据源的预设权重；另一种方式，是通过模型训练的方式，比如 LR 方式来训练得到最终的数据源预设权重，举例来说，首先给每个数据源赋个初始权重值，然后根据初选的小规模正负样本，将每个数据源作为其 $feature$ 来训练，最终迭代收敛后，模型即可吐出每个数据源的预设权重。

数据源预设权重确定后，再根据以下公式给每个初始标识打分：

$$score_i = \sum_{source=0}^n weight_{source} * exp^{-ln2*time} * sigmoid(action_{cnt})$$

其中， $source$ 代表的是数据源，这里有 n 个数据源； $weight$ 代表的是每个数据源上的预设权重； $time$ 为时间信息，在本是示例中，以 $abs(\text{用户行为发生时间}-\text{当前挖掘时间})$ ，即行为时间差的绝对值为例，其作为用户行为时间衰减参数，即行为发生距离当前时间越近，则其分值越大，距离当前时间越远，分值越小； $action$ 为频次信息，用于代表用户标识的行为频次，这里取了 $sigmoid$ 函数，对其做了归一化处理，其表示行为频次越多，分值越高。

5) 正负样本选择模块，被设置为根据用户打分模块中对初选人群的打分排序，选择排在前 N 位的标识 (N 值为多少可根据要挖掘的定向标识的不同，以及特征参数在标识中的数量分布，而自由设定)，选定后，前 N 位的标识即为正样本，在大盘活跃用户的标识中排除正样本集合，从剩余集合中选择同正样本 1:1 规模的人群作为负样本标识。

通过文本语义特征挖掘，获取训练数据正负样本，可以有效规避通常意义的种子人群规模过小，而导致模型训练特征不明显的问题；同时由于通过历史效果检验以及用户行为打分，可以用来衡量样本优劣，从而提升了样本选择的准确度。

根据本发明实施例，还提供了一种用于实施上述标识的获取方法的电子装置，如图 11 所示，该电子装置可以包括：一个或多个（图中仅示出一个）处理器 201、存储器 203、以及传输装置 205，如图 11 所示，该电子装置还可以包括输入输出设备 207。

其中，存储器 203 可用于存储计算机程序以及模块，如本发明实施例中的标识的获取方法和装置对应的程序指令/模块，处理器 201 被设置为通过运行存储在存储器 203 内的软件程序以及模块，从而执行各种功能应用以及数据处理，即实现上述的数据加载方法。存储器 203 可包括高速随机存储器，还可以包括非易失性存储器，如一个或者多个磁性存储装置、闪存、或者其他非易失性固态存储器。在一些实例中，存储器 203 可进一步包括相对于处理器 201 远程设置的存储器，这些远程存储器可以通过网络连接至终端。上述网络的实例包括但不限于互联网、企业内部网、局域网、移动通信网及其组合。

上述的传输装置 205 用于经由一个网络接收或者发送数据，还可以用于处理器与存储器之间的数据传输。上述的网络具体实例可包括有线网络及无线网络。在一个实例中，传输装置 205 包括一个网络适配器（Network Interface Controller，NIC），其可通过网线与其他网络设备与路由器相连从而可与互联网或局域网进行通讯。在一个实例中，传输装置 205 为射频（Radio Frequency，RF）模块，其用于通过无线方式与互联网进行通讯。

其中，可选地，存储器 203 用于存储应用程序。

处理器 201 可以通过传输装置 205 调用存储器 203 存储的应用程序，以执行下述步骤：从多个数据源中获取与预定操作对应的标识，其中，在所述多个数据源包括的目标数据源中记录有与所述标识对应的帐号和所述帐号执行过的所述预定操作；根据所述标识的特征信息以及预设特征词从所述标识中获取初始标识，其中，所述特征信息用于表示所述预定操作的特征；根据预设权重以及所述特征信息确定所述初始标识的特征参数，其中，所述预设权重与所述目标数据源对应，所述预设权重用于指示所述

目标数据源中的帐号执行所述预定操作的频率,所述特征参数用于指示所述初始标识执行所述预定操作的频率;从所述初始标识中获取第一目标标识,其中,所述第一目标标识是所述初始标识中所述特征参数高于预设参数的标识的集合。

处理器 201 还用于执行下述步骤:获取第一特征词与第二特征词,其中,所述预设特征词包括所述第一特征词和所述第二特征词;从所述标识中获取所述初始标识,其中,所述初始标识对应的特征信息中携带所述第一特征词且未携带所述第二特征词。

处理器 201 还用于执行下述步骤:获取所述预设权重,其中,所述预设权重的值越大表示所述目标数据源中的帐号执行所述预定操作的频率越高;从所述特征信息中获取时间信息和频次信息,其中,所述时间信息用于指示所述标识执行所述预定操作的时间,所述频次信息用于指示所述标识执行所述预定操作的频次;根据所述预设权重、所述时间信息以及所述频次信息确定所述特征参数,其中,所述特征参数的值越大表示所述初始标识执行所述预定操作的频率越高。

处理器 201 还用于执行下述步骤之一:获取所述目标数据源中执行所述预定操作的帐号在所述目标数据源中包括的全部帐号中所占的比例;根据所述比例为所述目标数据源分配所述预设权重,其中,所述比例越大的数据源分配的所述预设权重越大;获取第一标识集合与预设标识集合中相同标识的数量,其中,所述第一标识集合是所述初始标识中在一个所述目标数据源中包括的标识的集合;根据所述数量与所述第一标识集合中标识的数量之间的比值为所述目标数据源分配所述预设权重,其中,所述比值越大的数据源分配的所述预设权重越大。

处理器 201 还用于执行下述步骤:计算所述初始标识在每个所述目标数据源中对应的所述时间信息和所述频次信息的乘积;根据所述预设权重计算所述乘积的加权和,得到所述特征参数。

处理器 201 还用于执行下述步骤:从所述标识对应的所述预定操作中

获取用于表示所述预定操作的特征的信息，其中，所述用于表示所述预定操作的特征的信息包括：所述预定操作对应的特征词，所述时间信息和所述频次信息；将所述特征词、所述时间信息以及所述频次信息存储为预设格式，得到所述特征信息。

处理器 201 还用于执行下述步骤之一：将所述初始标识按照所述特征参数从高到低进行排列；从排列后的标识中选择出所述第一目标标识，其中，所述第一目标标识包括在排列后的标识中排在前 N 位的标识；从所述初始标识中获取所述特征参数的值大于或者等于预设值的所述第一目标标识。

处理器 201 还用于执行下述步骤：将所述第一目标标识与预设目标标识进行匹配；在所述第一目标标识与所述预设目标标识匹配成功的情况下，确定出所述第一目标标识为所需的标识；在所述第一目标标识与所述预设目标标识匹配不成功的情况下，重新获取所述第一目标标识。

处理器 201 还用于执行下述步骤：判断所述第一目标标识与所述预设目标标识中是否包括大于或者等于预设数量的相同标识；在判断出所述第一目标标识与所述预设目标标识中包括大于或者等于预设数量的相同标识的情况下，确定所述第一目标标识与所述预设目标标识匹配成功。

处理器 201 还用于执行下述步骤：获取所述多个数据源中包括的帐号对应的标识；从所述多个数据源中包括的帐号对应的标识中随机获取除所述第一目标标识之外的标识，得到第二目标标识，其中，所述第二目标标识中包括的标识的数量与所述第一目标标识中包括的标识的数量相同。

处理器 201 还用于执行下述步骤：根据所述第一目标标识和所述第二目标标识训练预测模型；根据所述预测模型从所述多个数据源包括的标识中为待推送资源获取待推送标识；向所述待推送标识推送所述待推送资源。

采用本发明实施例，提供了一种标识的获取方法的方案。通过从多个数据源中获取与预定操作对应的标识，其中，在多个数据源包括的目标数

据源中记录有与标识对应的帐号和帐号执行过的预定操作；根据标识的特征信息以及预设特征词从标识中获取初始标识，其中，特征信息用于表示预定操作的特征；根据预设权重以及特征信息确定初始标识的特征参数，其中，预设权重与目标数据源对应，预设权重用于指示目标数据源中的帐号执行预定操作的频率，特征参数用于指示初始标识执行预定操作的频率；从初始标识中获取第一目标标识，其中，第一目标标识是初始标识中特征参数高于预设参数的标识的集合。也就是说，在目标数据源中记录了标识对应的帐号以及帐号执行过的预定操作，从中获取预定操作对应的标识，使得标识的获取途径更加的广泛，避免了从单一的用户日志获取标识规模较小导致的获取的标识有偏的问题，再根据标识的特征信息以及预设特征词初步地筛选出初始标识，并根据预设权重和特征信息为初始标识确定特征参数来表示出初始标识执行该预定操作的频率，然后从初始标识中获取特征参数高于预设参数的第一目标标志，使得第一目标标识中包括的标识均为执行预定操作频率较高的标识，从而提高了获取用于训练的标识的准确度，进而克服相关技术中获取用于训练的标识的准确度低的问题。

可选地，本实施例中的具体示例可以参考上述实施例中所描述的示例，本实施例在此不再赘述。

本领域普通技术人员可以理解，图 11 所示的结构仅为示意，电子装置可以是智能手机（如 Android 手机、iOS 手机等）、平板电脑、掌上电脑以及移动互联网设备（Mobile Internet Devices, MID）、PAD 等终端设备。图 11 其并不对上述电子装置的结构造成限定。例如，电子装置还可包括比图 11 中所示更多或者更少的组件（如网络接口、显示装置等），或者具有与图 11 所示不同的配置。

本领域普通技术人员可以理解上述实施例的各种方法中的全部或部分步骤是可以通程序来指令终端设备相关的硬件来完成，该程序可以存储于一计算机可读存储介质中，存储介质可以包括：闪存盘、只读存储器（Read-Only Memory, ROM）、随机存取器（Random Access Memory, RAM）、

磁盘或光盘等。

本发明的实施例还提供了一种存储介质。可选地，在本实施例中，上述存储介质可以位于网络中的多个网络设备中的至少一个网络设备。

可选地，在本实施例中，存储介质被设置为存储用于执行以下步骤的程序代码：

S1，从多个数据源中获取与预定操作对应的标识，其中，在多个数据源包括的目标数据源中记录有与标识对应的帐号和帐号执行过的预定操作；

S2，根据标识的特征信息以及预设特征词从标识中获取初始标识，其中，特征信息用于表示预定操作的特征；

S3，根据预设权重以及特征信息确定初始标识的特征参数，其中，预设权重与目标数据源对应，预设权重用于指示目标数据源中的帐号执行预定操作的频率，特征参数用于指示初始标识执行预定操作的频率；

S4，从初始标识中获取第一目标标识，其中，第一目标标识是初始标识中特征参数高于预设参数的标识的集合。

可选地，存储介质还被设置为存储用于执行以下步骤的程序代码：

S1，获取第一特征词与第二特征词，其中，预设特征词包括第一特征词和第二特征词；

S2，从标识中获取初始标识，其中，初始标识对应的特征信息中携带第一特征词且未携带第二特征词。

可选地，存储介质还被设置为存储用于执行以下步骤的程序代码：获取预设权重，其中，预设权重的值越大表示目标数据源中的帐号执行预定操作的频率越高；从特征信息中获取时间信息和频次信息，其中，时间信息用于指示标识执行预定操作的时间，频次信息用于指示标识执行预定操作的频次；根据预设权重、时间信息以及频次信息确定特征参数，其中，

特征参数的值越大表示初始标识执行预定操作的频率越高。

可选地，存储介质还被设置为存储用于执行以下步骤的程序代码：获取目标数据源中执行预定操作的帐号在目标数据源中包括的全部帐号中所占的比例；根据比例为目标数据源分配预设权重，其中，比例越大的数据源分配的预设权重越大；或者，获取第一标识集合与预设标识集合中相同标识的数量，其中，第一标识集合是初始标识中在一个目标数据源中包括的标识的集合；根据数量与第一标识集合中标识的数量之间的比值为目标数据源分配预设权重，其中，比值越大的数据源分配的预设权重越大。

可选地，存储介质还被设置为存储用于执行以下步骤的程序代码：计算初始标识在每个目标数据源中对应的时间信息和频次信息的乘积；根据预设权重计算乘积的加权和，得到特征参数。

可选地，存储介质还被设置为存储用于执行以下步骤的程序代码：从标识对应的预定操作中获取用于表示预定操作的特征的信息，其中，用于表示预定操作的特征的信息包括：预定操作对应的特征词，时间信息和频次信息；将特征词、时间信息以及频次信息存储为预设格式，得到特征信息。

可选地，存储介质还被设置为存储用于执行以下步骤的程序代码：将初始标识按照特征参数从高到低进行排列；从排列后的标识中选择出第一目标标识，其中，第一目标标识包括在排列后的标识中排在前 N 位的标识；或者，从初始标识中获取特征参数的值大于或者等于预设值的第一目标标识。

可选地，存储介质还被设置为存储用于执行以下步骤的程序代码：将第一目标标识与预设目标标识进行匹配；在第一目标标识与预设目标标识匹配成功的情况下，确定出第一目标标识为所需的标识；在第一目标标识与预设目标标识匹配不成功的情况下，重新获取第一目标标识。

可选地，存储介质还被设置为存储用于执行以下步骤的程序代码：判

断第一目标标识与预设目标标识中是否包括大于或者等于预设数量的相同标识；在判断出第一目标标识与预设目标标识中包括大于或者等于预设数量的相同标识的情况下，确定第一目标标识与预设目标标识匹配成功。

可选地，存储介质还被设置为存储用于执行以下步骤的程序代码：获取多个数据源中包括的帐号对应的标识；从多个数据源中包括的帐号对应的标识中随机获取除第一目标标识之外的标识，得到第二目标标识，其中，第二目标标识中包括的标识的数量与第一目标标识中包括的标识的数量相同。

可选地，存储介质还被设置为存储用于执行以下步骤的程序代码：根据第一目标标识和第二目标标识训练预测模型；根据预测模型从多个数据源包括的标识中为待推送资源获取待推送标识；向待推送标识推送待推送资源。

可选地，在本实施例中，上述存储介质可以包括但不限于：U 盘、只读存储器（ROM，Read-Only Memory）、随机存取存储器（RAM，Random Access Memory）、移动硬盘、磁碟或者光盘等各种可以存储程序代码的介质。

可选地，本实施例中的具体示例可以参考上述实施例 1 和实施例 2 中所描述的示例，本实施例在此不再赘述。

上述本发明实施例序号仅仅为了描述，不代表实施例的优劣。

上述实施例中的集成的单元如果以软件功能单元的形式实现并作为独立的产品销售或使用，可以存储在上述计算机可读的存储介质中。基于这样的理解，本发明的技术方案本质上或者说对相关技术做出贡献的部分或者该技术方案的全部或部分可以以软件产品的形式体现出来，该计算机软件产品存储在存储介质中，包括若干指令用以使得一台或多台计算机设备（可为个人计算机、服务器或者网络设备）执行本发明各个实施

例所述方法的全部或部分步骤。

在本发明的上述实施例中，对各个实施例的描述都各有侧重，某个实施例中沒有详述的部分，可以参见其他实施例的相关描述。

在本申请所提供的几个实施例中，应该理解到，所揭露的客户端，可通过其它的方式实现。其中，以上所描述的装置实施例仅仅是示意性的，例如所述单元的划分，仅仅为一种逻辑功能划分，实际实现时可以有另外的划分方式，例如多个单元或组件可以结合或者可以集成到另一个系统，或一些特征可以忽略，或不执行。另一点，所显示或讨论的相互之间的耦合或直接耦合或通信连接可以是通过一些接口，单元或模块的间接耦合或通信连接，可以是电性或其它的形式。

所述作为分离部件说明的单元可以是或者也可以不是物理上分开的，作为单元显示的部件可以是或者也可以不是物理单元，即可以位于一个地方，或者也可以分布到多个网络单元上。可以根据实际的需要选择其中的部分或者全部单元来实现本实施例方案的目的。

另外，在本发明各个实施例中的各功能单元可以集成在一个处理单元中，也可以是各个单元单独物理存在，也可以两个或两个以上单元集成在一个单元中。上述集成的单元既可以采用硬件的形式实现，也可以采用软件功能单元的形式实现。

以上所述仅是本发明的优选实施方式，应当指出，对于本技术领域的普通技术人员来说，在不脱离本发明原理的前提下，还可以做出若干改进和润饰，这些改进和润饰也应视为本发明的保护范围。

工业实用性：

通过上述描述可知，本发明在目标数据源中记录了标识对应的帐号以及帐号执行过的预定操作，从中获取预定操作对应的标识，使得标识的获取途径更加的广泛，避免了从单一的用户日志获取标识规模较小导致的获取的标识有偏的问题，再根据标识的特征信息以及预设特征词初步地筛选

出初始标识,并根据预设权重和特征信息为初始标识确定特征参数来表示出初始标识执行该预定操作的频率,然后从初始标识中获取特征参数高于预设参数的第一目标标志,使得第一目标标识中包括的标识均为执行预定操作频率较高的标识,从而提高了获取用于训练的标识的准确度,进而克服相关技术中获取用于训练的标识的准确度低的问题。

权 利 要 求 书

1. 一种标识的获取方法，包括：

从多个数据源中获取与预定操作对应的标识，其中，在所述多个数据源包括的目标数据源中记录有与所述标识对应的帐号和所述帐号执行过的所述预定操作；

根据所述标识的特征信息以及预设特征词从所述标识中获取初始标识，其中，所述特征信息用于表示所述预定操作的特征；

根据预设权重以及所述特征信息确定所述初始标识的特征参数，其中，所述预设权重与所述目标数据源对应，所述预设权重用于指示所述目标数据源中的帐号执行所述预定操作的频率，所述特征参数用于指示所述初始标识执行所述预定操作的频率；

从所述初始标识中获取第一目标标识，其中，所述第一目标标识是所述初始标识中所述特征参数高于预设参数的标识的集合。

2. 根据权利要求 1 所述的方法，其中，根据所述标识的所述特征信息以及所述预设特征词从所述标识中获取所述初始标识包括：

获取第一特征词与第二特征词，其中，所述预设特征词包括所述第一特征词和所述第二特征词；

从所述标识中获取所述初始标识，其中，所述初始标识对应的特征信息中携带所述第一特征词且未携带所述第二特征词。

3. 根据权利要求 1 所述的方法，其中，根据所述预设权重以及所述特征信息确定所述初始标识的所述特征参数包括：

获取所述预设权重，其中，所述预设权重的值越大表示所述目标数据源中的帐号执行所述预定操作的频率越高；

从所述特征信息中获取时间信息和频次信息，其中，所述时间信息用于指示所述标识执行所述预定操作的时间，所述频次信息用于指示所述标识执行所述预定操作的频次；

根据所述预设权重、所述时间信息以及所述频次信息确定所述特征参数，其中，所述特征参数的值越大表示所述初始标识执行所述预定操作的频率越高。

4. 根据权利要求3所述的方法，其中，获取所述预设权重包括以下之一：

获取所述目标数据源中执行所述预定操作的帐号在所述目标数据源中包括的全部帐号中所占的比例；根据所述比例为所述目标数据源分配所述预设权重，其中，所述比例越大的数据源分配的所述预设权重越大；

获取第一标识集合与预设标识集合中相同标识的数量，其中，所述第一标识集合是所述初始标识中在一个所述目标数据源中包括的标识的集合；根据所述数量与所述第一标识集合中标识的数量之间的比值为所述目标数据源分配所述预设权重，其中，所述比值越大的数据源分配的所述预设权重越大。

5. 根据权利要求3所述的方法，其中，根据所述预设权重、所述时间信息以及所述频次信息确定所述特征参数包括：

计算所述初始标识在每个所述目标数据源中对应的所述时间信息和所述频次信息的乘积；

根据所述预设权重计算所述乘积的加权和，得到所述特征参数。

6. 根据权利要求3所述的方法，其中，在根据所述标识的特征信息以及预设特征词从所述标识中获取初始标识之前，所述方法还包括：

从所述标识对应的所述预定操作中获取用于表示所述预定操作的特征的信息，其中，所述用于表示所述预定操作的特征的信息包括：所述预定操作对应的特征词，所述时间信息和所述频次信息；

将所述特征词、所述时间信息以及所述频次信息存储为预设格式，得到所述特征信息。

7. 根据权利要求1所述的方法，其中，从所述初始标识中获取所述第一目标标识包括以下之一：

将所述初始标识按照所述特征参数从高到低进行排列；从排列后的标识中选择出所述第一目标标识，其中，所述第一目标标识包括在排列后的标识中排在前 N 位的标识；

从所述初始标识中获取所述特征参数的值大于或者等于预设值的所述第一目标标识。

8. 根据权利要求 1 至 7 中任一项所述的方法，其中，在从所述初始标识中获取所述第一目标标识之后，所述方法还包括：

将所述第一目标标识与预设目标标识进行匹配；

在所述第一目标标识与所述预设目标标识匹配成功的情况下，确定出所述第一目标标识为所需的标识；在所述第一目标标识与所述预设目标标识匹配不成功的情况下，重新获取所述第一目标标识。

9. 根据权利要求 8 所述的方法，其中，将所述第一目标标识与所述预设目标标识进行匹配包括：

判断所述第一目标标识与所述预设目标标识中是否包括大于或者等于预设数量的相同标识；

在判断出所述第一目标标识与所述预设目标标识中包括大于或者等于预设数量的相同标识的情况下，确定所述第一目标标识与所述预设目标标识匹配成功。

10. 根据权利要求 1 至 7 中任一项所述的方法，其中，在从所述初始标识中获取所述第一目标标识之后，所述方法还包括：

获取所述多个数据源中包括的帐号对应的标识；

从所述多个数据源中包括的帐号对应的标识中随机获取除所述第一目标标识之外的标识，得到第二目标标识，其中，所述第二目标标识中包括的标识的数量与所述第一目标标识中包括的标识的数量相同。

11. 根据权利要求 10 所述的方法，其中，在从所述多个数据源中包括的帐号对应的标识中随机获取除所述第一目标标识之外的标识，得到第

二目标标识之后，所述方法还包括：

根据所述第一目标标识和所述第二目标标识训练预测模型；

根据所述预测模型从所述多个数据源包括的标识中为待推送资源获取待推送标识；

向所述待推送标识推送所述待推送资源。

12. 一种标识的获取装置，包括一个或多个处理器，以及一个或多个存储程序单元的存储器，其中，所述程序单元由所述处理器执行，所述程序单元包括：

第一获取模块，被设置为从多个数据源中获取与预定操作对应的标识，其中，在所述多个数据源包括的目标数据源中记录有与所述标识对应的帐号和所述帐号执行过的操作；

第二获取模块，被设置为根据所述标识的特征信息以及预设特征词从所述标识中获取初始标识，其中，所述特征信息用于表示所述预定操作的特征；

确定模块，被设置为根据预设权重以及所述特征信息确定所述初始标识的特征参数，其中，所述预设权重与所述目标数据源对应，所述预设权重用于指示所述目标数据源中的帐号执行所述预定操作的频率，所述特征参数用于指示所述初始标识执行所述预定操作的频率；

第三获取模块，被设置为从所述初始标识中获取第一目标标识，其中，所述第一目标标识是所述初始标识中所述特征参数高于预设参数的标识的集合。

13. 根据权利要求 12 所述的装置，其中，所述第二获取模块包括：

第一获取单元，被设置为获取第一特征词与第二特征词，其中，所述预设特征词包括所述第一特征词和所述第二特征词；

第二获取单元，被设置为从所述标识中获取所述初始标识，其中，所述初始标识对应的特征信息中携带所述第一特征词且未携带所述第二特征词。

14. 根据权利要求 12 所述的装置，其中，所述确定模块包括：

第三获取单元，被设置为获取所述预设权重，其中，所述预设权重的值越大表示所述目标数据源中的帐号执行所述预定操作的频率越高；

第四获取单元，被设置为从所述特征信息中获取时间信息和频次信息，其中，所述时间信息用于指示所述标识执行所述预定操作的时间，所述频次信息用于指示所述标识执行所述预定操作的频次；

确定单元，被设置为根据所述预设权重、所述时间信息以及所述频次信息确定所述特征参数，其中，所述特征参数的值越大表示所述初始标识执行所述预定操作的频率越高。

15. 根据权利要求 12 所述的装置，其中，所述第三获取模块包括以下之一：

处理单元，被设置为将所述初始标识按照所述特征参数从高到低进行排列；从排列后的标识中选择出所述第一目标标识，其中，所述第一目标标识包括在排列后的标识中排在前 N 位的标识；

第五获取单元，被设置为从所述初始标识中获取所述特征参数的值大于或者等于预设值的所述第一目标标识。

16. 一种存储介质，其中，所述存储介质中存储有计算机程序，所述计算机程序被设置为运行时执行所述权利要求 1 至 11 任一项中所述的方法。

17. 一种电子装置，包括存储器和处理器，其中，所述存储器中存储有计算机程序，所述处理器被设置为通过所述计算机程序执行所述权利要求 1 至 11 任一项中所述的方法。

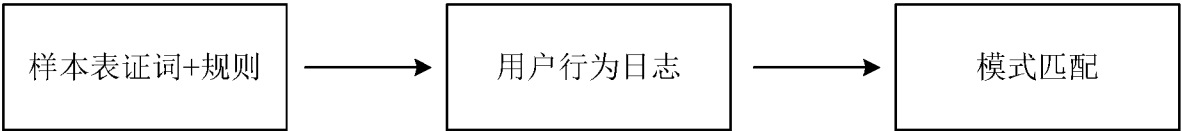


图 1

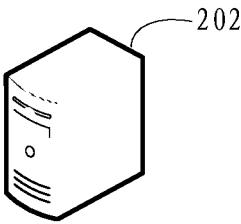


图 2

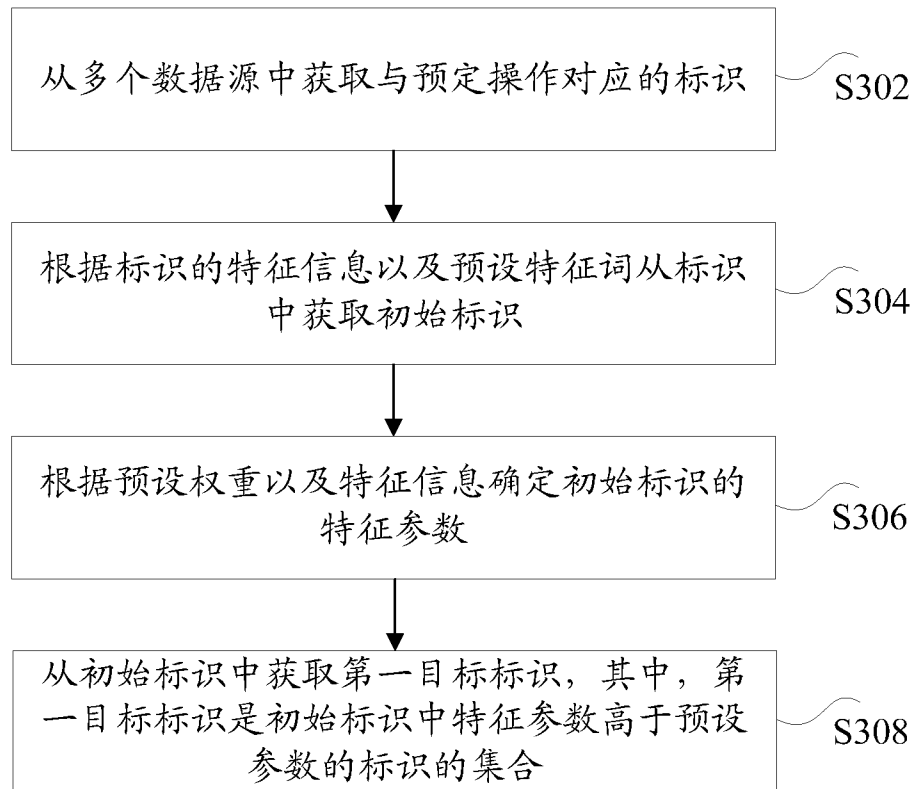


图 3

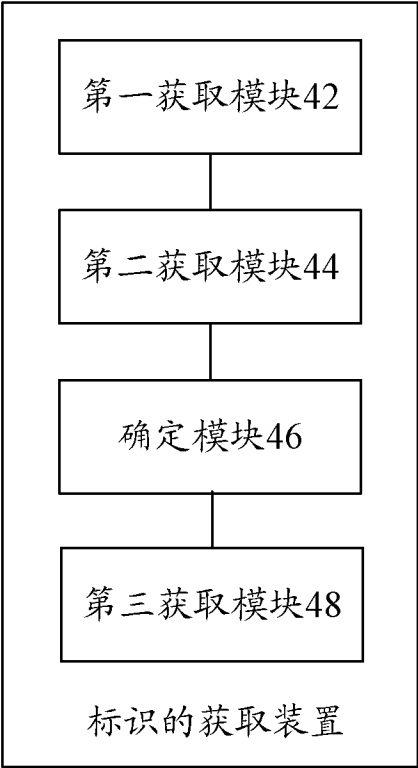


图 4

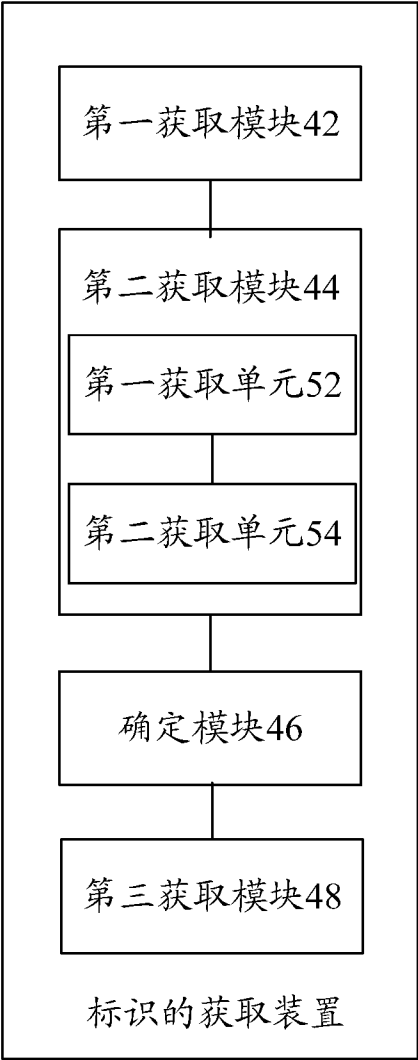


图 5

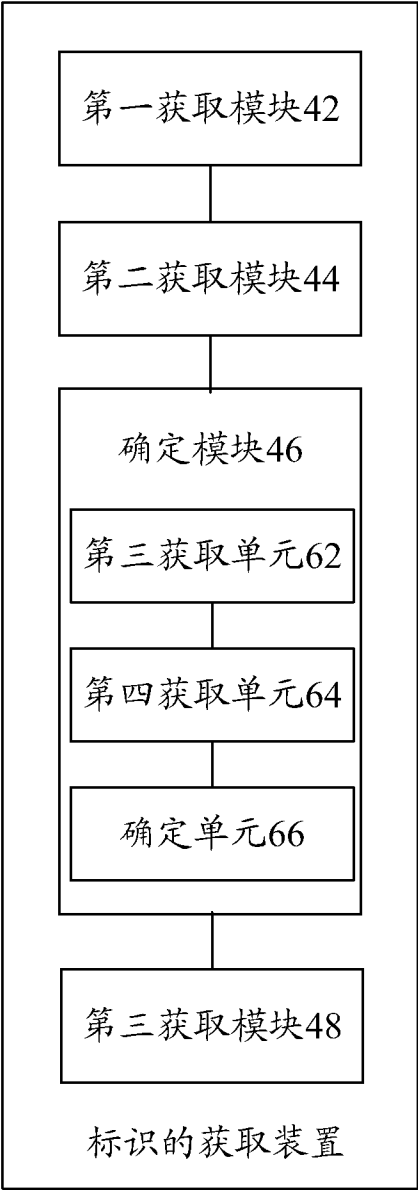


图 6

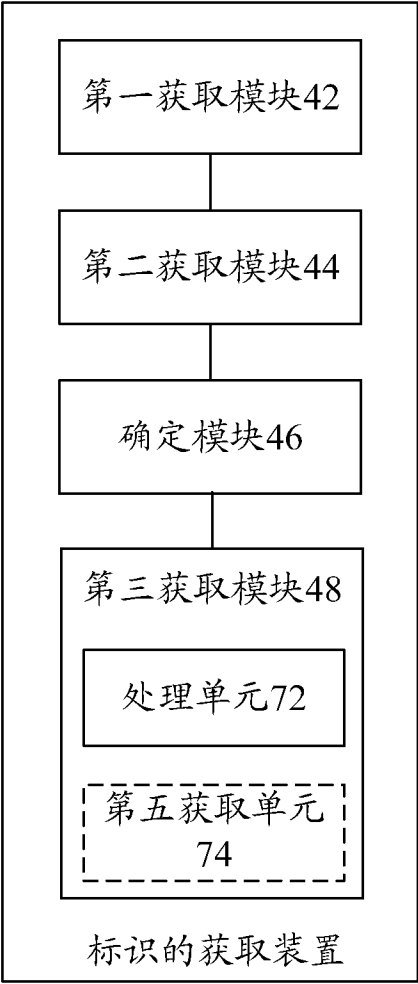


图 7

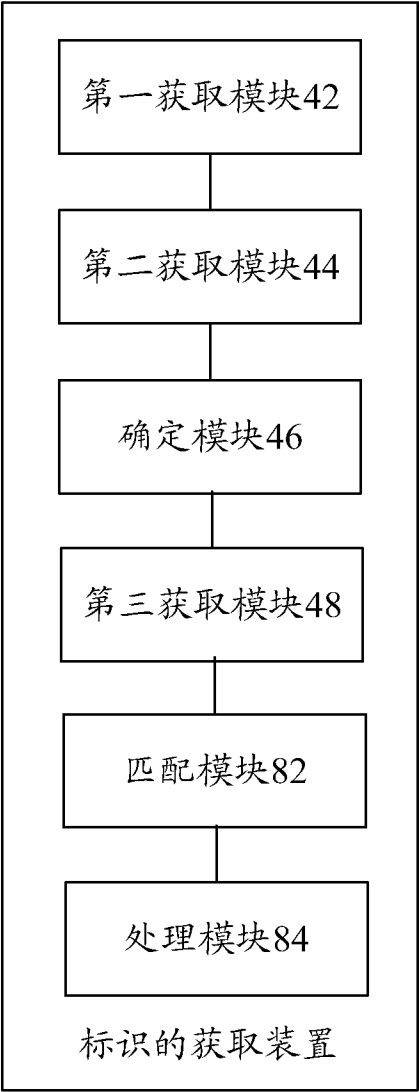


图 8

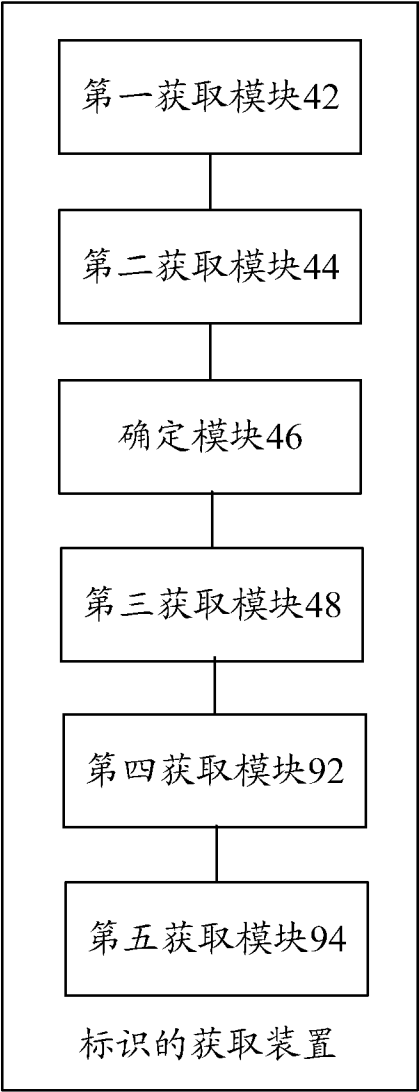


图 9

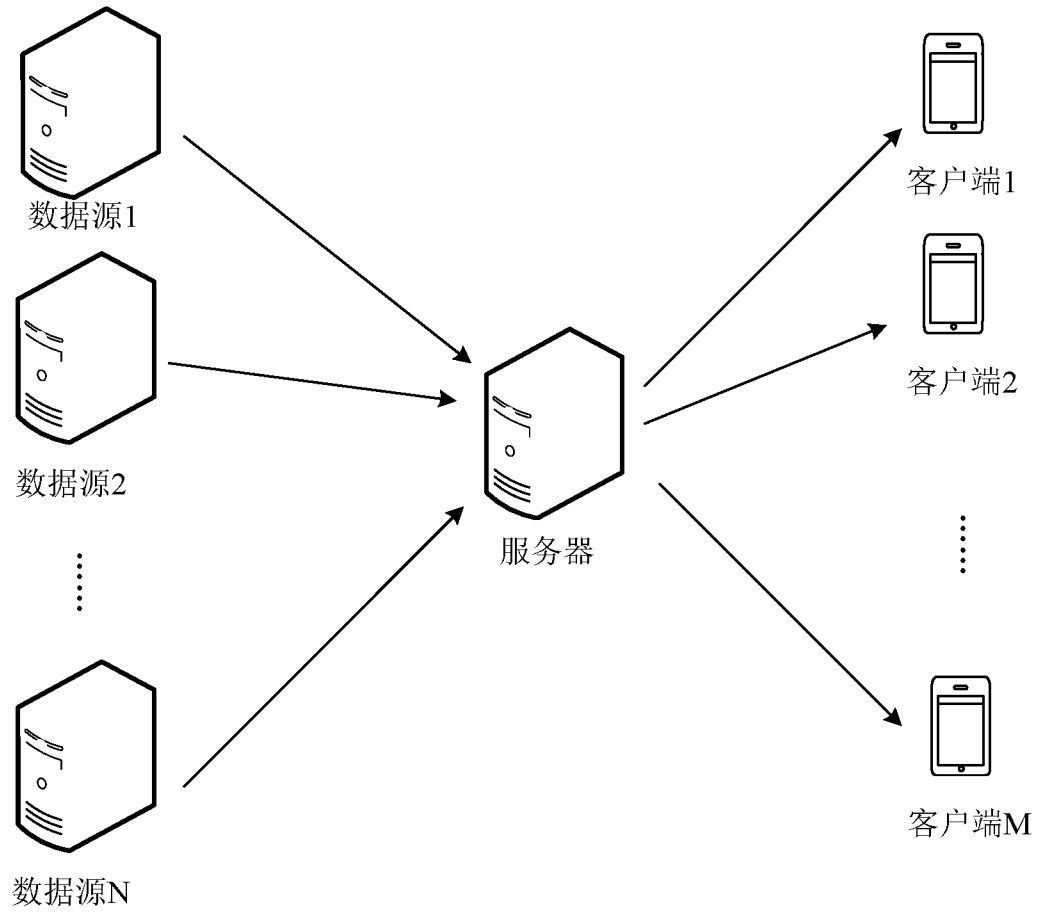


图 10

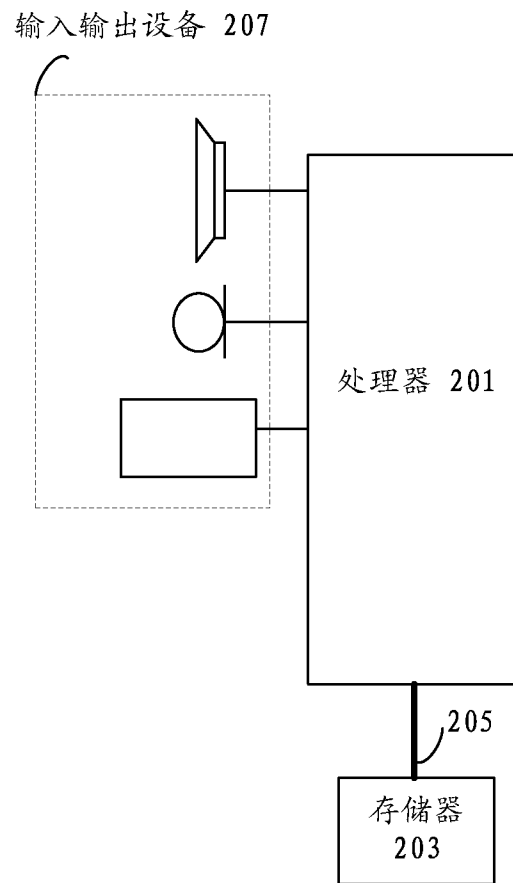


图 11

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2018/081337

A. CLASSIFICATION OF SUBJECT MATTER

G06F 17/00 (2006.01) i

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

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

CNABS, SIPOABS, CNKI, IEEE, CNTXT, DWPI, VEN, PATENTICS: 集合, 特征词, 记录, 特征, 权重, 历史, 操作, 账号, 关键词, 频率, 对应, 标识, set, feature, record, keyword, history, weight, operation, frequency, match, identifier

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	CN 106126592 A (BEIJING XIAOMI MOBILE SOFTWARE CO., LTD.), 16 November 2016 (16.11.2016), claims 1-13, description, paragraphs 85 and 89, and figures 1-7	1-17
A	CN 102831234 A (BEIJING UNIVERSITY OF POSTS AND TELECOMMUNICATIONS; BEIJING GOZAP TECHNOLOGY CO., LTD.), 19 December 2012 (19.12.2012), entire document	1-17
A	CN 104317865 A (NANJING UNIVERSITY OF POSTS AND TELECOMMUNICATIONS), 28 January 2015 (28.01.2015), entire document	1-17
A	CN 105430504 A (SHENZHEN INSTITUTES OF ADVANCED TECHNOLOGY, CHINESE ACADEMY OF SCIENCES), 23 March 2016 (23.03.2016), entire document	1-17
A	CN 102819804 A (ALIBABA GROUP HOLDING LIMITED), 12 December 2012 (12.12.2012), entire document	1-17

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

☒ See patent family annex.

* Special categories of cited documents:	"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"A" document defining the general state of the art which is not considered to be of particular relevance	"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"E" earlier application or patent but published on or after the international filing date	"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)	"&" document member of the same patent family
"O" document referring to an oral disclosure, use, exhibition or other means	
"P" document published prior to the international filing date but later than the priority date claimed	

Date of the actual completion of the international search 23 May 2018	Date of mailing of the international search report 31 May 2018
Name and mailing address of the ISA State Intellectual Property Office of the P. R. China No. 6, Xitucheng Road, Jimenqiao Haidian District, Beijing 100088, China Facsimile No. (86-10) 62019451	Authorized officer ZHANG, Guihua Telephone No. 86-(10)-62089416

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2018/081337

Patent Documents referred in the Report	Publication Date	Patent Family	Publication Date
CN 106126592 A	16 November 2016	None	
CN 102831234 A	19 December 2012	CN 102831234 B	22 April 2015
CN 104317865 A	28 January 2015	CN 104317865 B	09 June 2017
CN 105430504 A	23 March 2016	None	
CN 102819804 A	12 December 2012	WO 2012170475 A2	13 December 2012
		EP 2721563 A2	23 April 2014
		US 2012316960 A1	13 December 2012
		TW 201250619 A	16 December 2012
		WO 2012170475 A3	08 May 2014
		EP 2721563 A4	17 June 2015
		JP 2014519661 A	14 August 2014
		TW I529642 B	11 April 2016
		JP 5818980 B2	18 November 2015

国际检索报告

国际申请号

PCT/CN2018/081337

A. 主题的分类

G06F 17/00(2006.01)i

按照国际专利分类(IPC)或者同时按照国家分类和IPC两种分类

B. 检索领域

检索的最低限度文献(标明分类系统和分类号)

G06F

包含在检索领域中的除最低限度文献以外的检索文献

在国际检索时查阅的电子数据库(数据库的名称, 和使用的检索词(如使用))

CNABS, SIPOABS, CNKI, IEEE, CNTXT, DWPI, VEN, PATENTICS:集合, 特征词, 记录, 特征, 权重, 历史, 操作, 账号, 关键词, 频率, 对应, 标识, set, feature, record, keyword, history, weight, operation, frequency, match, identifier

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 106126592 A (京小米移动软件有限公) 2016年 11月 16日 (2016 - 11 - 16) 权利要求1-13, 说明书第85、89段, 图1-7	1-17
A	CN 102831234 A (北京邮电大学 北京格致璞科技有限公司) 2012年 12月 19日 (2012 - 12 - 19) 全文	1-17
A	CN 104317865 A (南京邮电大学) 2015年 1月 28日 (2015 - 01 - 28) 全文	1-17
A	CN 105430504 A (中国科学院深圳先进技术研究院) 2016年 3月 23日 (2016 - 03 - 23) 全文	1-17
A	CN 102819804 A (阿里巴巴集团控股有限公司) 2012年 12月 12日 (2012 - 12 - 12) 全文	1-17

☐ 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

2018年 5月 23日

国际检索报告邮寄日期

2018年 5月 31日

ISA/CN的名称和邮寄地址

中华人民共和国国家知识产权局(ISA/CN)
中国北京市海淀区蓟门桥西土城路6号 100088

传真号 (86-10)62019451

受权官员

张桂华

电话号码 86-(10)-62089416

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2018/081337

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	106126592	A	2016年 11月 16日	无			
CN	102831234	A	2012年 12月 19日	CN	102831234	B	2015年 4月 22日
CN	104317865	A	2015年 1月 28日	CN	104317865	B	2017年 6月 9日
CN	105430504	A	2016年 3月 23日	无			
CN	102819804	A	2012年 12月 12日	WO	2012170475	A2	2012年 12月 13日
				EP	2721563	A2	2014年 4月 23日
				US	2012316960	A1	2012年 12月 13日
				TW	201250619	A	2012年 12月 16日
				WO	2012170475	A3	2014年 5月 8日
				EP	2721563	A4	2015年 6月 17日
				JP	2014519661	A	2014年 8月 14日
				TW	I529642	B	2016年 4月 11日
				JP	5818980	B2	2015年 11月 18日

表 PCT/ISA/210 (同族专利附件) (2015年1月)