(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
2022 年 2 月 3 日 (03.02.2022)

WIPO | PCT

(10) 国际公布号
WO 2022/021977 A1

(51) 国际专利分类号:
G06F 16/9535 (2019.01)

(21) 国际申请号: PCT/CN2021/090947

(22) 国际申请日: 2021 年 4 月 29 日 (29.04.2021)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
202010763020.X 2020年7月31日 (31.07.2020) CN

(71) 申请人: 平安科技（深圳）有限公司 (PING AN TECHNOLOGY (SHENZHEN) CO., LTD.) [CN/CN]; 中国广东省深圳市福田区福田街道福安社区益田路5033号平安金融中心23楼, Guangdong 518000 (CN)。

(72) 发明人: 孙家棣 (SUN, Jiadi); 中国广东省深圳市福田区福田街道福安社区益田路5033号平安金融中心23楼, Guangdong 518000 (CN)。
马宁 (MA, Ning); 中国广东省深圳市福田区福田街道福安社区益田路5033号平安金融中心23楼, Guangdong 518000 (CN)。

(74) 代理人: 深圳市隆天联鼎知识产权代理有限公司 (SHENZHEN LUNGTIN LIANDING INTELLECTUAL PROPERTY AGENT LTD.); 中国广东省深圳市福田区南园路上田大厦4A刘抗美, Guangdong 518000 (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, IT, 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:** UNDERGROUND INDUSTRY ACCOUNT DETECTION METHOD AND APPARATUS, COMPUTER DEVICE, AND MEDIUM

(54) **发明名称:** 黑产账号检测方法、装置、计算机设备和介质

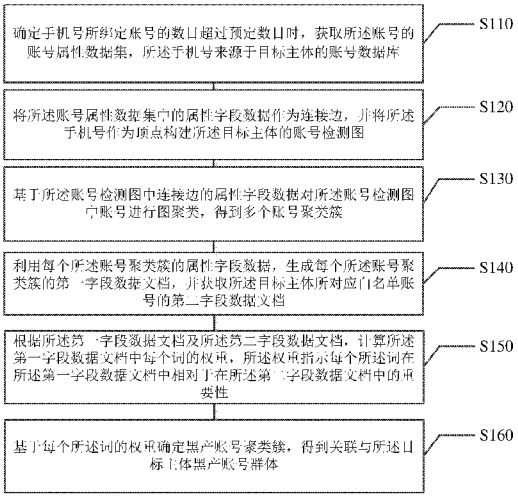

图 1

S110 Upon determining that the number of accounts bound to a mobile phone number of a user exceeds a predetermined number, obtain the account attribute data set of an account, wherein the mobile phone number is from an account database of a target subject

S120 Use attribute field data in the account attribute data set as a connection edge, and use the mobile phone number as a vertex to construct an account detection graph of the target subject

S130 Perform graph clustering on the accounts in the account detection graph according to the attribute field data of the connection edge in the account detection graph to obtain a plurality of account clusters

S140 Use the attribute field data of each account cluster to generate a first field data document of each account cluster, and obtain a second field data document of a whitelist account corresponding to the target subject

S150 Calculate the weight of each word in the first field data document according to the first field data document and the second field data document, wherein the weight indicates the importance of each word in the first field data document with respect to the word in the second field data document

S160 Determine an underground industry account cluster according to the weight of each word so as to obtain an underground industry account group associated with the target subject

(57) **Abstract:** An artificial intelligence-based underground industry account detection method and a related apparatus. The method comprises: upon determining that the number of accounts bound to a mobile phone number of a user exceeds a predetermined number, obtaining the account attribute data set of an account, wherein the user is associated with a target subject (S110); using attribute field data in the account attribute data set as a connection edge, and using the mobile phone number as a vertex to construct an account detection graph of the target subject (S120); performing graph clustering on the accounts in the account detection graph to obtain a plurality of account clusters (S130); using the attribute field data of each account cluster to generate a first field data document, and obtaining a second field data document of a whitelist account corresponding to the target subject (S140); calculating the weight of each word in the first field data document (S150); and determining an underground industry account cluster according to the weight of each word (S160). The solution further relates to the field of blockchains, and the account attribute data set can be stored in a blockchain, thereby effectively improving the accuracy of underground industry account detection.

WO 2022/021977 A1

[见续页]

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

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

本国际公布:

— 包括国际检索报告(条约第21条(3))。

(57) 摘要：一种基于人工智能的黑产账号检测方法及相关装置，该方法包括：确定用户的手机号所绑定账号的数目超过预定数目时，获取所述账号的账号属性数据集，所述用户关联于所述目标主体（S110）；将所述账号属性数据集中的属性字段数据作为连接边，并将所述手机号作为顶点构建所述目标主体的账号检测图（S120）；对账号检测图中账号进行图聚类得到多个账号聚类簇（S130）；利用每个账号聚类簇的属性字段数据，生成第一字段数据文档，并获取目标主体所对应白名单账号的第二字段数据文档（S140）；计算第一字段数据文档中每个词的权重（S150），以基于每个词的权重确定黑产账号聚类簇（S160）。该方案还涉及区块链领域，账号属性数据集可存储于区块链，有效提升黑产账号检测的准确性。

黑产账号检测方法、装置、计算机设备和介质

本申请要求于 2020 年 7 月 31 日提交中国专利局、申请号为 CN 202010763020.X，发明名称为“基于人工智能的黑产账号检测方法及相关装置”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

本申请涉及人工智能技术领域，特别是涉及一种基于人工智能的黑产账号检测方法、装置、计算机设备和计算机可读存储介质。

背景技术

一般地，随着互联网大众化应用越来越广泛，正常用户在享受网络便利的同时，也伴随着网络黑产带来的危险。目前，网络黑产已经规模化、链条化，黑产形式越来越多样，防作弊难度越来越大。黑产已逐利为目的，有需求，就有市场。

随着技术的进步，黑产攻击已经成为各大公司非常重视的问题，无时无刻不在面临着黑产的攻击。黑产无论是如何变现，都需要先注册大量的虚假账号，以量攻击。

发明人意识到，目前，业务风险识别需要打击黑产的行为，识别打击虚假黑产账号。业内目前主要是通过专家经验规则来识别和打击虚假账号。专家规则识别面比较单一，比较窄，主要是定向精准识别和打击，因为逻辑较简单，容易被黑产行为识别和绕过。

发明内容

在人工智能技术领域，为了解决上述技术问题，本申请的目的在于提供一种基于人工智能的黑产账号检测方法、装置、计算机设备和计算机可读存储介质。

第一方面，提供了一种基于人工智能的黑产账号检测方法，包括：

确定手机号所绑定账号的数目超过预定数目时，获取所述账号的账号属性数据集，所述手机号来源于目标主体的账号数据库；

将所述账号属性数据集中的属性字段数据作为连接边，并将所述手机号作为顶点构建所述目标主体的账号检测图；

基于所述账号检测图中连接边的属性字段数据对所述账号检测图中账号进行图聚类，得到多个账号聚类簇；

利用每个所述账号聚类簇的属性字段数据，生成每个所述账号聚类簇的第一字段数据文档，并获取所述目标主体所对应白名单账号的第二字段数据文档；

根据所述第一字段数据文档及所述第二字段数据文档，计算所述第一字段数据文档中每个词的权重，所述权重指示每个所述词在所述第一字段数据文档中相对于在所述第二字段数据文档中的重要性；

基于每个所述词的权重确定黑产账号聚类簇，得到关联与所述目标主体黑产账号群体。

第二方面，提供了一种基于人工智能的黑产账号检测装置，包括：

获取模块，用于确定手机号所绑定账号的数目超过预定数目时，获取所述账号的账号属性数据集，所述手机号来源于目标主体的账号数据库；

构建模块，用于将所述账号属性数据集中的属性字段数据作为连接边，并将所述手机号作为顶点构建所述目标主体的账号检测图；

聚类模块，用于基于所述账号检测图中连接边的属性字段数据对所述账号检测图中账号进行图聚类，得到多个账号聚类簇；

生成模块，用于利用每个所述账号聚类簇的属性字段数据，生成每个所述账号聚类簇的第一字段数据文档，并获取所述目标主体所对应白名单账号的第二字段数据文档；

计算模块，用于根据所述第一字段数据文档及所述第二字段数据文档，计算所述第

一字段数据文档中每个词的权重，所述权重指示每个所述词在所述第一字段数据文档中相对于在所述第二字段数据文档中的重要性；

确定模块，用于基于每个所述词的权重确定黑产账号聚类簇，得到关联与所述目标主体黑产账号群体。

第三方面，提供了一种计算机设备，包括存储器和处理器，所述存储器用于存储所述处理器的基于人工智能的黑产账号检测的程序，所述处理器配置为经由执行所述基于人工智能的黑产账号检测的程序来执行以下处理：确定手机号所绑定账号的数目超过预定数目时，获取所述账号的账号属性数据集，所述手机号来源于目标主体的账号数据库；将所述账号属性数据集中的属性字段数据作为连接边，并将所述手机号作为顶点构建所述目标主体的账号检测图；基于所述账号检测图中连接边的属性字段数据对所述账号检测图中账号进行图聚类，得到多个账号聚类簇；利用每个所述账号聚类簇的属性字段数据，生成每个所述账号聚类簇的第一字段数据文档，并获取所述目标主体所对应白名单账号的第二字段数据文档；根据所述第一字段数据文档及所述第二字段数据文档，计算所述第一字段数据文档中每个词的权重，所述权重指示每个所述词在所述第一字段数据文档中相对于在所述第二字段数据文档中的重要性；基于每个所述词的权重确定黑产账号聚类簇，得到关联与所述目标主体黑产账号群体。

第四方面，提供了一种存储有计算机可读指令的计算机可读存储介质，其上存储有基于人工智能的黑产账号检测的程序，所述基于人工智能的黑产账号检测的程序被处理器执行时实现以下处理：确定手机号所绑定账号的数目超过预定数目时，获取所述账号的账号属性数据集，所述手机号来源于目标主体的账号数据库；将所述账号属性数据集中的属性字段数据作为连接边，并将所述手机号作为顶点构建所述目标主体的账号检测图；基于所述账号检测图中连接边的属性字段数据对所述账号检测图中账号进行图聚类，得到多个账号聚类簇；利用每个所述账号聚类簇的属性字段数据，生成每个所述账号聚类簇的第一字段数据文档，并获取所述目标主体所对应白名单账号的第二字段数据文档；根据所述第一字段数据文档及所述第二字段数据文档，计算所述第一字段数据文档中每个词的权重，所述权重指示每个所述词在所述第一字段数据文档中相对于在所述第二字段数据文档中的重要性；基于每个所述词的权重确定黑产账号聚类簇，得到关联与所述目标主体黑产账号群体。

上述基于人工智能的黑产账号检测方法、装置、计算机设备和计算机可读存储介质，首先，确定来源于目标主体的账号数据库的手机号所绑定账号的数目超过预定数目时，获取账号的账号属性数据集；对目标主体中账号进行初步筛选，排除手机号所绑定账号小于预定数目的账号，得到待检测的账号属性数据集，缩小检测范围同时提升检测可靠性；然后，将账号属性数据集中的属性字段数据作为连接边，并将手机号作为顶点构建所述目标主体的账号检测图；基于账号检测图中连接边的属性字段数据对账号检测图中账号进行图聚类，得到多个账号聚类簇；通过将关联账号的手机号作为顶点构建图，然后基于属性字段数据图聚类得到账号聚类簇，可靠聚类得到账号团伙；然后，利用每个账号聚类簇的属性字段数据，生成每个账号聚类簇的第一字段数据文档，并获取目标主体所对应白名单账号的第二字段数据文档；可以便于基于数据文档进行数据分析，同时通过正常的第二字段数据文档作为第一字段数据文档的对照，保证黑产账号检测的准确性；最后，根据第一字段数据文档及第二字段数据文档，计算第一字段数据文档中每个词的权重，以基于每个所述词的权重确定黑产账号聚类簇，该权重指示每个词在所述第一字段数据文档中相对于在所述第二字段数据文档中的重要性。可以通过每个词在所述第一字段数据文档中相对于在所述第二字段数据文档中的重要性，准确判断账号聚类簇是否黑产账号团伙。

应当理解的是，以上的一般描述和后文的细节描述仅是示例性的，并不能限制本申

请。

附图说明

图1示意性示出一种基于人工智能的黑产账号检测方法的流程图。

图2示意性示出一种基于人工智能的黑产账号检测方法的应用场景示例图。

图3示意性示出一种获取账号的账号属性数据集的方法流程图。

图4示意性示出一种基于人工智能的黑产账号检测装置的方框图。

图5示意性示出一种用于实现上述基于人工智能的黑产账号检测方法的计算机设备的示例框图。

图6示意性示出一种用于实现上述基于人工智能的黑产账号检测方法的计算机可读存储介质。

具体实施方式

现在将参考附图更全面地描述示例实施方式。然而，示例实施方式能够以多种形式实施，且不应被理解为限于在此阐述的范例；相反，提供这些实施方式使得本申请将更加全面和完整，并将示例实施方式的构思全面地传达给本领域的技术人员。所描述的特征、结构或特性可以以任何合适的方式结合在一个或更多实施方式中。在下面的描述中，提供许多具体细节从而给出对本申请的实施方式的充分理解。然而，本领域技术人员将意识到，可以实践本申请的技术方案而省略所述特定细节中的一个或更多，或者可以采用其它的方法、组元、装置、步骤等。在其它情况下，不详细示出或描述公知技术方案以避免喧宾夺主而使得本申请的各方面变得模糊。

此外，附图仅为本申请的示意性图解，并非一定是按比例绘制。图中相同的附图标记表示相同或类似的部分，因而将省略对它们的重复描述。附图中所示的一些方框图是功能实体，不一定必须与物理或逻辑上独立的实体相对应。可以采用软件形式来实现这些功能实体，或在一个或多个硬件模块或集成电路中实现这些功能实体，或在不同网络和/或处理器装置和/或微控制器装置中实现这些功能实体。

本示例实施方式中首先提供了基于人工智能的黑产账号检测方法，该基于人工智能的黑产账号检测方法可以运行于服务器，也可以运行于服务器集群或云服务器等，当然，本领域技术人员也可以根据需求在其他平台运行本申请的方法，本示例性实施例中对此不做特殊限定。参考图 1 所示，该基于人工智能的黑产账号检测方法可以包括以下步骤：

步骤 S110，确定手机号所绑定账号的数目超过预定数目时，获取所述账号的账号属性数据集，所述手机号来源于目标主体的账号数据库；

步骤 S120，将所述账号属性数据集中的属性字段数据作为连接边，并将所述手机号作为顶点构建所述目标主体的账号检测图；

步骤 S130，基于所述账号检测图中连接边的属性字段数据对所述账号检测图中账号进行图聚类，得到多个账号聚类簇；

步骤 S140，利用每个所述账号聚类簇的属性字段数据，生成每个所述账号聚类簇的第一字段数据文档，并获取所述目标主体所对应白名单账号的第二字段数据文档；

步骤 S150，根据所述第一字段数据文档及所述第二字段数据文档，计算所述第一字段数据文档中每个词的权重，所述权重指示每个所述词在所述第一字段数据文档中相对于在所述第二字段数据文档中的重要性；

步骤 S160，基于每个所述词的权重确定黑产账号聚类簇，得到关联与所述目标主体黑产账号群体。

上述基于人工智能的黑产账号检测方法中，首先，确定来源于目标主体的账号数据库的手机号所绑定账号的数目超过预定数目时，获取账号的账号属性数据集；对目标主体中账号进行初步筛选，排除手机号所绑定账号小于预定数目的账号，得到待检测的账

号属性数据集，缩小检测范围同时提升检测可靠性。

然后，将账号属性数据集中的属性字段数据作为连接边，并将手机号作为顶点构建所述目标主体的账号检测图；基于账号检测图中连接边的属性字段数据对账号检测图中账号进行图聚类，得到多个账号聚类簇；通过将关联账号的手机号作为顶点构建图，然后基于属性字段数据图聚类得到账号聚类簇，可靠聚类得到账号团伙。

然后，利用每个账号聚类簇的属性字段数据，生成每个账号聚类簇的第一字段数据文档，并获取目标主体所对应白名单账号的第二字段数据文档；可以便于基于数据文档进行数据分析，同时通过正常的第二字段数据文档作为第一字段数据文档的对照，保证黑产账号检测的准确性。

最后，根据第一字段数据文档及第二字段数据文档，计算第一字段数据文档中每个词的权重，以基于每个所述词的权重确定黑产账号聚类簇，该权重指示每个词在所述第一字段数据文档中相对于在所述第二字段数据文档中的重要性。可以通过每个词在所述第一字段数据文档中相对于在所述第二字段数据文档中的重要性，可靠判断账号聚类簇是否黑产账号团伙。

下面，将结合附图对本示例实施方式中上述基于人工智能的黑产账号检测方法中的各步骤进行详细的解释以及说明。

在步骤 S110，确定手机号所绑定账号的数目超过预定数目时，获取所述账号的账号属性数据集，所述手机号来源于目标主体的账号数据库。

在本示例的实施方式中，参考图 2 所示，服务器 210 可以从服务器 220 中获取目标主体所关联账号的账号属性数据集；然后，服务器 210 可以确定用户的手机号所绑定账号的数目超过预定数目时，获取绑定账号的数目超过预定数目的手机号所对应的所有账号的账号属性数据集。其中，服务器 210、服务器 220 可以是电脑、手机等各种具有指令处理功能、数据存储功能的终端设备，在此不做特殊限定。

本示例的事实方式中，服务器 210 与服务器 220 为区块链中的节点服务器，进而基于区块链中数据不可更改以及安全性，服务器 210 可以安全可靠地从服务器 220 中获取到目标主体所关联账号的账号属性数据集。

每个账号的账号属性数据集中包括账号相关属性字段的字段数据，可以包括手机号、设备、网络环境、登录密码等相关属性字段的字段数据，例如，账号密码、手机号及登录设备 id 等。目标主体可以是任意的企业或者平台等。

用户的手机号所绑定账号的数目超过预定数目时，说明用户的手机号绑定的账号具有黑产行为嫌疑，其中，预定数目可以根据实际情况设定，为预设的一个手机号所关联的账号的标准数量，一个手机号所关联的账号超过该阈值说明具有黑产嫌疑，例如可以是 5 个等。

一示例中，可以通过应用端采集到的网络环境、设备参数、注册密码等账号属性数据。网络黑产可能伪装网络环境、设备参数、注册密码等维度，但注册绑定同一用户手机号的账号指标需求是不可能绕过的。例如，目标主体对于用户手机号的注册绑定账号要求是每个月底生效时，可以先根据绑定同一用户手机号一个月内不小于目标主体绑定账号所设定的预定数目，进行捞取账号属性数据集。

确定用户的手机号所绑定账号的数目超过预定数目时，获取所述账号的账号属性数据集，可以对目标主体关联的所有账号进行初步筛选，排除手机号所绑定账号小于预定数目的账号，得到剩余的待检测的账号属性数据集，缩小检测范围的同时提升检测准确性。

一种实施例中，参考图 3 所示，确定手机号所绑定账号的数目超过预定数目时，获取所述账号的账号属性数据集包括：

步骤 S310，获取所述目标主体与所述手机号的业务关联条件，所述业务关联条件指

示所述手机号在目标业务中可以绑定账号的数目阈值，所述目标业务来源于所述目标主体；

步骤 S320，当绑定了所述手机号的账号超过所述数目阈值时，获取所述账号的账号属性数据集。

业务关联条件指示用户的手机号在目标业务中可以绑定账号的数目阈值，即在目标主体举行的某次业务活动中设定的手机号可以绑定账号的数目阈值，与目标业务相适应，实现根据不同的业务精准监控黑产嫌疑账号。

步骤 S120，将所述账号属性数据集中的属性字段数据作为连接边，并将所述手机号作为顶点构建所述目标主体的账号检测图。

在本示例的实施方式中，将账号属性数据集中的属性字段数据作为连接边及将手机号作为顶点构建检测图，即将账号关联的手机号作为顶点，各账号之间根据字段数据的关联关系，将字段作为连接边，连接各关联账号，得到检测图，可以包含获取的账号之间的各种关联关系。

一种实施例中，将所述账号属性数据集中的属性字段数据作为连接边，并将所述手机号作为顶点构建所述目标主体的账号检测图，包括：

获取所述账号属性数据集中的指纹型字段和类别型字段，所述指纹型字段至少包括登录设备标识、登录密码和登录设备开机时间，所述类别型字段至少包括登录设备机型、系统版本、设备总存储空间、登录网络地址、无线网卡的物理地址；

将第一预定数目个所述指纹型字段的组合及第二预定数目个所述类别型字段的组合在所述账号属性数据集中的字段数据组合作为连接边，并将所述字段数据组合对应的手机号作为顶点构建账号检测图。

将字段定义成两类：指纹型字段和类别型字段。对于指纹型字段，任意第一预定数目个字段数据放在一起可作为检测图的连接边；而类别型字段需要第二预定数目个字段数据、放在一起作为检测图的连接边。

一种实施例中，所述第一预定数目为 2，所述第二预定数目大于等于 3 且小于等于 5。

指纹型字段单独一个字段作为连接边进行筛选，也可以两个组合作为边在一起，可以有效避免误伤碰撞的情况。例如，黑产改某个指纹型字段的字段数据碰巧和正常账号的一样了，两个放在一起组合作为连接边使用，减少了误伤碰撞的概率。同样，类别型字段多个放在一起也是更加精准的筛选数据。

例如，（a）Ios 系统指纹型变量是登录设备标识 id、登录密码和登录设备开机时间 boottime。单个字段数据与手机号个数的对应关系如（a1-a3）：

（a1）设备 id 个数和手机号个数对应关系是 1：1.06。（a2）登录密码个数和手机号个数关系是 1：1.51。（a3）boottime 个数和手机号个数关系是 1：1.18。

而如（a4-a6）中两两组合起来，两个字段数据组合和手机号个数几乎是一对一的关系：（a4）设备 id 和 boottime 放在一起和手机号个数关系是 1：1.04。（a5）登录密码和 boottime 放在一起和手机号个数关系是 1：1.01。（a6）设备 id 和登录密码放在一起和手机号个数关系是 1：1.02。

例如，类别型变量包括登录设备机型、系统版本、设备总存储空间、登录网络地址 ip、无线网卡的物理地址 wifimac 等，通常，除了上述指纹型变量以外，都可以归结为类别型变量。单个字段数据与手机号个数的对应关系如（b1-b2）：（b1）机型个数和手机号个数关系是 1：28470.36，且机型数目总数通常为 70 种。（b2）设备总存储空间个数和手机号个数关系是 1：134.34。通过组合可以有效减少对应手机号个数。

步骤 S130，基于所述账号检测图中连接边的属性字段数据对所述账号检测图中账号进行图聚类，得到多个账号聚类簇。

在本示例的实施方式中，可以使用现有的图聚类方法对账号检测图进行图聚类，得

到账号聚类簇。这样可以基于构建账号检测图构建出账号的关系网，并基于属性字段数据进行账号的聚类，得到相似的账号聚类簇。

一种实施例中，基于所述账号检测图中连接边的属性字段数据对所述账号检测图中账号进行图聚类，得到多个账号聚类簇，包括：

基于所述连接边的属性字段数据，对所述账号检测图利用 Connected Component 算法进行图聚类处理，得到多个账号群体；

从所述多个账号群体中，获取包含手机号个数大于等于预定个数且关联于同一登录网络地址的账号群体，得到第一账号群体组合；

从所述多个账号群体中，获取包含手机号个数大于等于预定个数且关联于同一无线网卡的物理地址的账号群体，得到第二账号群体组合；

将所述第一账号群体组合及所述第二账号群体组合确定为所述账号聚类簇。

手机号为顶点，上述步骤中定义的连接边，使用 Connected Component 算法图聚类计算，得到多个节点簇。

Connected Components 算法，即连通体算法用一标识 id 标注图中每个连通体（多个账号群体），将连通体中序号最小的顶点的标识 id 作为连通体的标识 id。如果在图 G 中，任意 2 个顶点（手机号）之间都存在路径，那么称 G 为连通图，否则称该图为非连通图，则其中的极大连通子图称为连通体。

然后，再二次图聚类，以第一次聚类结果的群编号（标识 id）为顶点，首先，从所述多个账号群体中，获取包含手机号个数大于等于预定个数且关联于同一登录网络地址的账号群体，得到第一账号群体组合，例如，获取包含手机号个数大于等于 3 且关联于同一登录网络地址的账号群体，得到第一账号群体组合。然后，从所述多个账号群体中，获取包含手机号个数大于等于所述预定个数且关联于同一无线网卡的物理地址的账号群体，得到第二账号群体组合，例如，（获取包含手机号个数大于等于 3 且关联于同一无线网卡的物理地址的账号群体，得到第二账号群体组合。二次图聚类的使用主要应对于秒拨动态 ip（登录网络地址、无线网卡的物理地址）和将本应该是同一团伙的小群体合并。

例如，首先，当 A、B 及 C 三个账号群体基于 ip 连接；然后，当 A、D 及 E 三个群体连接；这样的话：A、B 及 C、D、E 则是共同算一个团伙账号聚类簇。

黑产会对 ip 进行伪装几个手机号换一次 ip 或者 wifimac，这样，在一次图聚类结果中，存在手机号个数比较少的群组，而群组中 ip 或 wifimac 是一样的，将这些群组 id 作为顶点，ip 或 wifimac 作为连接边实现二次聚类。

步骤 S140，利用每个所述账号聚类簇的属性字段数据，生成每个所述账号聚类簇的第一字段数据文档，并获取所述目标主体所对应白名单账号的第二字段数据文档。

在本示例的实施方式中，第一字段数据文档及第二字段数据文档可以是文本文档或者表格等。

目标主体所对应白名单账号，即白名单用户的账号，可以为目标主体对应的主体内部用户的账号属性数据集，例如，某个机构的员工的账号相关数据，可以确定为非黑产的数据。

目标主体所对应白名单账号的第二字段数据文档，可以通过的白名单账号的属性字段数据生成。

通过生成第一字段数据文档及第二字段数据文档可以便于基于数据文档进行数据分析，同时通过正常的第二字段数据文档作为第一字段数据文档的对照，保证黑产账号检测的准确性。

步骤 S150，根据所述第一字段数据文档及所述第二字段数据文档，计算所述第一字段数据文档中每个词的权重，所述权重指示每个所述词在所述第一字段数据文档中相对于在所述第二字段数据文档中的重要性。

在本示例的实施方式中，通过计算指示每个词在第一字段数据文档中相对于在第二字段数据文档中的重要性的权重，可以获取每个账号聚类簇的第一字段数据文档中权重“独特”的词（即属性字段数据），进而，账号团伙中存在该团伙“独特”的属性字段数据，大概率是黑产修改的模拟器参数。

一种实施例中，根据所述第一字段数据文档及所述第二字段数据文档，计算所述第一字段数据文档中每个词的权重，包括：

对于所述第一字段数据文档中的每个词，计算每个词在所述第一字段数据文档中出现的第一频率；

对于所述第一字段数据文档中的每个词，计算每个所述词在所述第一字段数据文档及所述第二字段数据文档中同时出现的第二频率；

将所述第一频率与所述第二频率的乘积，作为每个所述词的权重。

计算每个词在第一字段数据文档中出现的第一频率，可以得到每个词在待检测的第一字段数据文档中的重要性；然后，计算每个词在第一字段数据文档及第二字段数据文档中同时出现的第二频率，可以得到每个词的全局重要性，最后将第一频率与第二频率的乘积，作为每个词的权重，可以通过权重从全局数据集的角度指示每个词在第一字段数据文档中相对于在第二字段数据文档中的重要性。

一种实施例中，根据所述第一字段数据文档及所述第二字段数据文档，计算所述第一字段数据文档中每个词的权重，包括：

基于公式$tf-idf(t,d)=tf(t,d)\times idf(t)$，$idf(t)=\log(N/e)$计算每个词的权重，其中，所述$tf-idf(t,d)$为权重，所述$t$为词，所述$d$为所述第一字段数据文档，所述$tf(t,d)$为所述词在所述第一字段数据文档中出现的频数，所述$idf(t)$是文档中词的逆向文本频率，所述$N$为所述第一字段数据文档及所述第二字段数据文档的总数，e是所述第一字段数据文档及所述第二字段数据文档中出现所述词的文档数量。

通过公式$tf-idf(t,d)=tf(t,d)\times idf(t)$，$idf(t)=\log(N/e)$可以通过 TF-IDF 算法可以精准、高效识别黑产团伙账号（黑产账号聚类簇）的模拟器参数。团伙账号中存在 TF-IDF 权值较大的词，说明团伙账号中存在该团伙账号“独特”的词，大概率是模拟器参数。这样可以节省检测资源，同样的检测个数条件下，按照 TF-IDF 权重排序可以捞出更多的黑产团伙账号。实验证明，按照此标准排序，可以在相同的检测数量条件下，发现更多的黑产团伙账号。

步骤 S160，基于每个所述词的权重确定黑产账号聚类簇，得到关联与所述目标主体黑产账号群体。

通过每个词在所述第一字段数据文档中相对于在所述第二字段数据文档中的重要性，可靠判断账号聚类簇是否黑产账号团伙。

一种实施例中，基于每个所述词的权重确定黑产账号聚类簇，包括：

确定所述权重高于预定权值的词所来源的所述第一字段数据文档，作为黑产数据文档；

将所述黑产数据文档对应的账号聚类簇确定为黑产账号团伙。

预定权值可以根据实际情况设定。存在权重高于预定权值的词，说明权重高于预定权值的词所来源的第一字段数据文档中数据异常，确定为黑产数据文档，进而，可以，将黑产数据文档对应的账号聚类簇确定为黑产账号团伙。

一种实施例中，基于每个所述词的权重确定黑产账号聚类簇，包括：

计算每个所述第一字段数据文档中词的权重平均值；

确定所述权重平均值高于预定平均值的所述第一字段数据文档，作为黑产数据文档；

将所述黑产数据文档对应的账号聚类簇确定为黑产账号团伙。

计算每个第一字段数据文档中词的权重平均值，可以综合考虑所有词的权重，基于第一字段数据文档全局考虑账号聚类簇的异常情况。进而，确定权重平均值高于预定平均值的第一字段数据文档，作为黑产数据文档，可以从全局可靠检测出黑产账号团伙。

本申请还提供了一种基于人工智能的黑产账号检测装置。参考图 4 所示，该基于人工智能的黑产账号检测装置可以包括获取模块 410、构建模块 420、聚类模块 430、生成模块 440、计算模块 450 及确定模块 460。其中：

获取模块 410 可以用于确定用户的手机号所绑定账号的数目超过预定数目时，获取所述账号的账号属性数据集，所述用户关联于所述目标主体；

构建模块 420 可以用于将所述账号属性数据集中的属性字段数据作为连接边，并将所述手机号作为顶点构建所述目标主体的账号检测图；

聚类模块 430 可以用于基于所述账号检测图中连接边的属性字段数据对所述账号检测图中账号进行图聚类，得到多个账号聚类簇；

生成模块 440 可以用于利用每个所述账号聚类簇的属性字段数据，生成每个所述账号聚类簇的第一字段数据文档，并获取所述目标主体所对应白名单账号的第二字段数据文档；

计算模块 450 可以用于根据所述第一字段数据文档及所述第二字段数据文档，计算所述第一字段数据文档中每个词的权重，所述权重指示每个所述词在所述第一字段数据文档中相对于在所述第二字段数据文档中的重要性；

确定模块 460 可以用于基于每个所述词的权重确定黑产账号聚类簇，得到关联与所述目标主体黑产账号群体。

在一个实施例中，所述获取模块被进一步配置为：

获取目标主体与所述手机号的业务关联条件，所述业务关联条件指示所述手机号在目标业务中可以绑定账号的数目阈值，所述目标业务来源于所述目标主体；

当绑定了所述手机号的账号超过所述数目阈值时，获取所述账号的账号属性数据集。

在一个实施例中，所述聚类模块被进一步配置为：

基于所述连接边的属性字段数据，对所述账号检测图利用 Connected Component 算法进行图聚类处理，得到多个账号群体；

从所述多个账号群体中，获取包含手机号个数大于等于预定个数且关联于同一登录网络地址的账号群体，得到第一账号群体组合；

从所述多个账号群体中，获取包含手机号个数大于等于预定个数且关联于同一无线网卡的物理地址的账号群体，得到第二账号群体组合；

将所述第一账号群体组合及所述第二账号群体组合确定为所述账号聚类簇。

在一个实施例中，所述计算模块被进一步配置为：

对于所述第一字段数据文档中的每个词，计算每个词在所述第一字段数据文档中出现的第一频率；

对于所述第一字段数据文档中的每个词，计算每个所述词在所述第一字段数据文档及所述第二字段数据文档中同时出现的第二频率；

将所述第一频率与所述第二频率的乘积，作为每个所述词的权重。

在一个实施例中，所述计算模块被进一步配置为：

基于公式$tf-idf(t,d)=tf(t,d)\times idf(t)$，$idf(t)=\log(N/e)$计算每个词的权重，其中，所述$tf-idf(t,d)$为权重，所述$t$为词，所述$d$为所述第一字段数据文档，所述$tf(t,d)$为所述词在所述第一字段数据文档中出现的频数，所述$idf(t)$是文档中词的逆向文本频率，所述$N$为所述第一字段数据文档及所述第二字段数据文档的总数，e是所述第一字段数据文档及所述第二字段数据文档中出现所述词的文档数量。

在一个实施例中，所述确定模块被进一步配置为：

确定所述权重高于预定权值的词所来源的所述第一字段数据文档，作为黑产数据文档；

将所述黑产数据文档对应的账号聚类簇确定为黑产账号团伙。

在一个实施例中，所述确定模块被进一步配置为：

计算每个所述第一字段数据文档中词的权重平均值；

确定所述权重平均值高于预定平均值的所述第一字段数据文档，作为黑产数据文档；

将所述黑产数据文档对应的账号聚类簇确定为黑产账号团伙。

应当注意，尽管在上文详细描述中提及了用于动作执行的设备的若干模块或者单元，但是这种划分并非强制性的。实际上，根据本申请的实施方式，上文描述的两个或更多模块或者单元的特征和功能可以在一个模块或者单元中具体化。反之，上文描述的一个模块或者单元的特征和功能可以进一步划分为由多个模块或者单元来具体化。

此外，尽管在附图中以特定顺序描述了本申请中方法的各个步骤，但是，这并非要求或者暗示必须按照该特定顺序来执行这些步骤，或是必须执行全部所示的步骤才能实现期望的结果。附加的或备选的，可以省略某些步骤，将多个步骤合并为一个步骤执行，以及/或者将一个步骤分解为多个步骤执行等。

根据本申请的第三方面，还提供了一种计算机设备，执行上述任一所示的基于人工智能的黑产账号检测方法的全部或者部分步骤。该计算机设备包括：

至少一个处理器；以及

与所述至少一个处理器通信连接的存储器；其中，

所述存储器存储有可被所述至少一个处理器执行的指令，所述指令被所述至少一个处理器执行，以使所述至少一个处理器能够执行如上述任一个示例性实施例所示出的基于人工智能的黑产账号检测方法。

所属技术领域的技术人员能够理解，本申请的各个方面可以实现为系统、方法或程序产品。因此，本申请的各个方面可以具体实现为以下形式，即：完全的硬件实施方式、完全的软件实施方式（包括固件、微代码等），或硬件和软件方面结合的实施方式，这里可以统称为“电路”、“模块”或“系统”。

下面参照图 5 来描述根据本申请的这种实施方式的计算机设备 500。图 5 显示的计算机设备 500 仅仅是一个示例，不应对本申请实施例的功能和使用范围带来任何限制。

如图 5 所示，计算机设备 500 以通用计算设备的形式表现。计算机设备 500 的组件可以包括但不限于：上述至少一个处理单元 510、上述至少一个存储单元 520、连接不同系统组件（包括存储单元 520 和处理单元 510）的总线 530。

其中，所述存储单元存储有程序代码，所述程序代码可以被所述处理单元 510 执行，使得所述处理单元 510 执行本说明书上述“实施例方法”部分中描述的根据本申请各种示例性实施方式的步骤。例如，所述处理单元 510 可以执行如图 1 中所示的步骤 S110，确定手机号所绑定账号的数目超过预定数目时，获取所述账号的账号属性数据集，所述手机号来源于目标主体的账号数据库；步骤 S120，将所述账号属性数据集中的属性字段数据作为连接边，并将所述手机号作为顶点构建所述目标主体的账号检测图；步骤 S130，基于所述账号检测图中连接边的属性字段数据对所述账号检测图中账号进行图聚类，得到多个账号聚类簇；步骤 S140，利用每个所述账号聚类簇的属性字段数据，生成每个所述账号聚类簇的第一字段数据文档，并获取所述目标主体所对应白名单账号的第二字段数据文档；步骤 S150，根据所述第一字段数据文档及所述第二字段数据文档，计算所述第一字段数据文档中每个词的权重，所述权重指示每个所述词在所述第一字段数据文档中相对于在所述第二字段数据文档中的重要性；步骤 S160，基于每个所述词的权重确定黑产账号聚类簇，得到关联与所述目标主体黑产账号群体。

存储单元 520 可以包括易失性存储单元形式的可读介质，例如随机存取存储单元（RAM）5201 和/或高速缓存存储单元 5202，还可以进一步包括只读存储单元（ROM）5203。

存储单元 520 还可以包括具有一组（至少一个）程序模块 5205 的程序/实用工具 5204，这样的程序模块 5205 包括但不限于：操作系统、一个或者多个应用程序、其它程序模块以及程序数据，这些示例中的每一个或某种组合中可能包括网络环境的实现。

总线 530 可以为表示几类总线结构中的一种或多种，包括存储单元总线或者存储单元控制器、外围总线、图形加速端口、处理单元或者使用多种总线结构中的任意总线结构的局域总线。

计算机设备 500 也可以与一个或多个外部设备 700（例如键盘、指向设备、蓝牙设备等）通信，还可与一个或者多个使得用户能与该计算机设备 500 交互的设备通信，和/或与使得该计算机设备 500 能与一个或多个其它计算机设备进行通信的任何设备（例如路由器、调制解调器等等）通信。这种通信可以通过输入/输出（I/O）接口 550 进行，还可以包括与输入/输出（I/O）接口 550 连接的显示单元 540。并且，计算机设备 500 还可以通过网络适配器 560 与一个或者多个网络（例如局域网（LAN），广域网（WAN）和/或公共网络，例如因特网）通信。如图所示，网络适配器 560 通过总线 530 与计算机设备 500 的其它模块通信。应当明白，尽管图中未示出，可以结合计算机设备 500 使用其它硬件和/或软件模块，包括但不限于：微代码、设备驱动器、冗余处理单元、外部磁盘驱动阵列、RAID 系统、磁带驱动器以及数据备份存储系统等。

通过以上的实施方式的描述，本领域的技术人员易于理解，这里描述的示例实施方式可以通过软件实现，也可以通过软件结合必要的硬件的方式来实现。因此，根据本申请实施方式的技术方案可以以软件产品的形式体现出来，该软件产品可以存储在一个非易失性存储介质（可以是 CD-ROM，U 盘，移动硬盘等）中或网络上，包括若干指令以使得一台计算机设备（可以是个人计算机、服务器、终端装置、或者网络设备等）执行根据本申请实施方式的方法。

根据本申请的第四方面，还提供了一种计算机可读存储介质，其上存储有能够实现本说明书上述方法的程序产品，所述计算机可读存储介质可以是非易失性，也可以是易失性。在一些可能的实施方式中，本申请的各个方面还可以实现为一种程序产品的形式，其包括程序代码，当所述程序产品在终端设备上运行时，所述程序代码用于使所述终端设备执行本说明书上述“示例性方法”部分中描述的根据本申请各种示例性实施方式的步骤。

参考图 6 所示，描述了根据本申请的实施方式的用于实现上述方法的程序产品 600，其可以采用便携式紧凑盘只读存储器(CD-ROM)并包括程序代码，并可以在终端设备，例如个人电脑上运行。然而，本申请的程序产品不限于此，在本文件中，计算机可读存储介质可以是任何包含或存储程序的有形介质，该程序可以被指令执行系统、装置或者器件使用或者与其结合使用。

所述程序产品可以采用一个或多个可读介质的任意组合。可读介质可以是可读信号介质或者可读存储介质。可读存储介质例如可以为但不限于电、磁、光、电磁、红外线、或半导体的系统、装置或器件，或者任意以上的组合。可读存储介质的更具体的例子（非穷举的列表）包括：具有一个或多个导线的电连接、便携式盘、硬盘、随机存取存储器（RAM）、只读存储器（ROM）、可擦式可编程只读存储器（EPROM 或闪存）、光纤、便携式紧凑盘只读存储器(CD-ROM)、光存储器件、磁存储器件、或者上述的任意合适的组合。

计算机可读信号介质可以包括在基带中或者作为载波一部分传播的数据信号，其中承载了可读程序代码。这种传播的数据信号可以采用多种形式，包括但不限于电磁信号、

光信号或上述的任意合适的组合。可读信号介质还可以是可读存储介质以外的任何可读介质，该可读介质可以发送、传播或者传输用于由指令执行系统、装置或者器件使用或者与其结合使用的程序。

可读介质上包含的程序代码可以用任何适当的介质传输，包括但不限于无线、有线、光缆、RF 等等，或者上述的任意合适的组合。

可以以一种或多种程序设计语言的任意组合来编写用于执行本申请操作的程序代码，所述程序设计语言包括面向对象的程序设计语言—诸如 Java、C++等，还包括常规的过程式程序设计语言—诸如“C”语言或类似的程序设计语言。程序代码可以完全地在用户计算机设备上执行、部分地在用户计算机设备上执行、作为一个独立的软件包执行、部分在用户计算机设备上部分在远程计算机设备上执行、或者完全在远程计算机设备或服务器上执行。在涉及远程计算机设备的情形中，远程计算机设备可以通过任意种类的网络，包括局域网（LAN）或广域网（WAN），连接到用户计算机设备，或者，可以连接到外部计算机设备（例如利用因特网服务提供商来通过因特网连接）。

此外，上述附图仅是根据本申请示例性实施例的方法所包括的处理的示意性说明，而不是限制目的。易于理解，上述附图所示的处理并不表明或限制这些处理的时间顺序。另外，也易于理解，这些处理可以是例如在多个模块中同步或异步执行的。

应当理解的是，本申请并不局限于上面已经描述并在附图中示出的精确结构，并且可以在不脱离其范围执行各种修改和改变。本申请的范围仅由所附的权利要求来限制。

权利要求

1、一种基于人工智能的黑产账号检测方法，包括：

确定手机号所绑定账号的数目超过预定数目时，获取所述账号的账号属性数据集，所述手机号来源于目标主体的账号数据库；

将所述账号属性数据集中的属性字段数据作为连接边，并将所述手机号作为顶点构建所述目标主体的账号检测图；

基于所述账号检测图中连接边的属性字段数据对所述账号检测图中账号进行图聚类，得到多个账号聚类簇；

利用每个所述账号聚类簇的属性字段数据，生成每个所述账号聚类簇的第一字段数据文档，并获取所述目标主体所对应白名单账号的第二字段数据文档；

根据所述第一字段数据文档及所述第二字段数据文档，计算所述第一字段数据文档中每个词的权重，所述权重指示每个所述词在所述第一字段数据文档中相对于在所述第二字段数据文档中的重要性；

基于每个所述词的权重确定黑产账号聚类簇，得到关联与所述目标主体黑产账号群体。

2、根据权利要求 1 所述的方法，其中，所述确定手机号所绑定账号的数目超过预定数目时，获取所述账号的账号属性数据集，包括：

获取目标主体与所述手机号的业务关联条件，所述业务关联条件指示所述手机号在目标业务中可以绑定账号的数目阈值，所述目标业务来源于所述目标主体；

当绑定了所述手机号的账号超过所述数目阈值时，获取所述账号的账号属性数据集。

3、根据权利要求 1 所述的方法，其中，所述基于所述账号检测图中连接边的属性字段数据对所述账号检测图中账号进行图聚类，得到多个账号聚类簇，包括：

基于所述连接边的属性字段数据，对所述账号检测图利用 Connected Component 算法进行图聚类处理，得到多个账号群体；

从所述多个账号群体中，获取包含手机号个数大于等于预定个数且关联于同一登录网络地址的账号群体，得到第一账号群体组合；

从所述多个账号群体中，获取包含手机号个数大于等于预定个数且关联于同一无线网卡的物理地址的账号群体，得到第二账号群体组合；

将所述第一账号群体组合及所述第二账号群体组合确定为所述账号聚类簇。

4、根据权利要求 1 所述的方法，其中，所述根据所述第一字段数据文档及所述第二字段数据文档，计算所述第一字段数据文档中每个词的权重，包括：

对于所述第一字段数据文档中的每个词，计算每个词在所述第一字段数据文档中出现的第一频率；

对于所述第一字段数据文档中的每个词，计算每个所述词在所述第一字段数据文档及所述第二字段数据文档中同时出现的第二频率；

将所述第一频率与所述第二频率的乘积，作为每个所述词的权重。

5、根据权利要求 1 所述的方法，其中，所述根据所述第一字段数据文档及所述第二字段数据文档，计算所述第一字段数据文档中每个词的权重，包括：

基于公式$tf-idf(t,d)=tf(t,d)\times idf(t)$，$idf(t)=\log(N/e)$计算每个词的权重，其中，所述$tf-idf(t,d)$为权重，所述$t$为词，所述$d$为所述第一字段数据文档，所述$tf(t,d)$为所述词在所述第一字段数据文档中出现的频数，所述$idf(t)$是文档中词的逆向文本频率，所述$N$为所述第一字段数据文档及所述第二字段数据文档的总数，e 是所述第一字段数据文档及所述第二字段数据文档中出现所述词的文档数量。

6、根据权利要求 1 所述的方法，其中，所述基于每个所述词的权重确定黑产账号聚类簇，包括：

确定所述权重高于预定权值的词所来源的所述第一字段数据文档，作为黑产数据文档；

将所述黑产数据文档对应的账号聚类簇确定为黑产账号团伙。

7、根据权利要求 1 所述的方法，其中，所述基于每个所述词的权重确定黑产账号聚类簇，包括：

计算每个所述第一字段数据文档中词的权重平均值；

确定所述权重平均值高于预定平均值的所述第一字段数据文档，作为黑产数据文档；

将所述黑产数据文档对应的账号聚类簇确定为黑产账号团伙。

8、一种基于人工智能的黑产账号检测装置，包括：

获取模块，用于确定手机号所绑定账号的数目超过预定数目时，获取所述账号的账号属性数据集，所述手机号来源于目标主体的账号数据库；

构建模块，用于将所述账号属性数据集中的属性字段数据作为连接边，并将所述手机号作为顶点构建所述目标主体的账号检测图；

聚类模块，用于基于所述账号检测图中连接边的属性字段数据对所述账号检测图中账号进行图聚类，得到多个账号聚类簇；

生成模块，用于利用每个所述账号聚类簇的属性字段数据，生成每个所述账号聚类簇的第一字段数据文档，并获取所述目标主体所对应白名单账号的第二字段数据文档；

计算模块，用于根据所述第一字段数据文档及所述第二字段数据文档，计算所述第一字段数据文档中每个词的权重，所述权重指示每个所述词在所述第一字段数据文档中相对于在所述第二字段数据文档中的重要性；

确定模块，用于基于每个所述词的权重确定黑产账号聚类簇，得到关联与所述目标主体黑产账号群体。

9、一种计算机设备，包括存储器和处理器，所述存储器中存储有计算机可读指令，所述计算机可读指令被所述处理器执行时，使得所述处理器执行：

确定手机号所绑定账号的数目超过预定数目时，获取所述账号的账号属性数据集，所述手机号来源于目标主体的账号数据库；

将所述账号属性数据集中的属性字段数据作为连接边，并将所述手机号作为顶点构建所述目标主体的账号检测图；

基于所述账号检测图中连接边的属性字段数据对所述账号检测图中账号进行图聚类，得到多个账号聚类簇；

利用每个所述账号聚类簇的属性字段数据，生成每个所述账号聚类簇的第一字段数据文档，并获取所述目标主体所对应白名单账号的第二字段数据文档；

根据所述第一字段数据文档及所述第二字段数据文档，计算所述第一字段数据文档中每个词的权重，所述权重指示每个所述词在所述第一字段数据文档中相对于在所述第二字段数据文档中的重要性；

基于每个所述词的权重确定黑产账号聚类簇，得到关联与所述目标主体黑产账号群体。

10、根据权利要求 9 所述的计算机设备，其中，所述确定手机号所绑定账号的数目超过预定数目时，获取所述账号的账号属性数据集，包括：

获取目标主体与所述手机号的业务关联条件，所述业务关联条件指示所述手机号在目标业务中可以绑定账号的数目阈值，所述目标业务来源于所述目标主体；

当绑定了所述手机号的账号超过所述数目阈值时，获取所述账号的账号属性数据集。

11、根据权利要求 9 所述的计算机设备，其中，所述基于所述账号检测图中连接边的属性字段数据对所述账号检测图中账号进行图聚类，得到多个账号聚类簇，包括：

基于所述连接边的属性字段数据，对所述账号检测图利用 Connected Component 算法进

行图聚类处理，得到多个账号群体；

从所述多个账号群体中，获取包含手机号个数大于等于预定个数且关联于同一登录网络地址的账号群体，得到第一账号群体组合；

从所述多个账号群体中，获取包含手机号个数大于等于预定个数且关联于同一无线网卡的物理地址的账号群体，得到第二账号群体组合；

将所述第一账号群体组合及所述第二账号群体组合确定为所述账号聚类簇。

12、根据权利要求 9 所述的计算机设备，其中，所述根据所述第一字段数据文档及所述第二字段数据文档，计算所述第一字段数据文档中每个词的权重，包括：

对于所述第一字段数据文档中的每个词，计算每个词在所述第一字段数据文档中出现的第一频率；

对于所述第一字段数据文档中的每个词，计算每个所述词在所述第一字段数据文档及所述第二字段数据文档中同时出现的第二频率；

将所述第一频率与所述第二频率的乘积，作为每个所述词的权重。

13、根据权利要求 9 所述的计算机设备，其中，所述根据所述第一字段数据文档及所述第二字段数据文档，计算所述第一字段数据文档中每个词的权重，包括：

基于公式$tf-idf(t,d)=tf(t,d)\times idf(t)$，$idf(t)=\log(N/e)$计算每个词的权重，其中，所述$tf-idf(t,d)$为权重，所述$t$为词，所述$d$为所述第一字段数据文档，所述$tf(t,d)$为所述词在所述第一字段数据文档中出现的频数，所述$idf(t)$是文档中词的逆向文本频率，所述$N$为所述第一字段数据文档及所述第二字段数据文档的总数，e 是所述第一字段数据文档及所述第二字段数据文档中出现所述词的文档数量。

14、根据权利要求 9 所述的计算机设备，其中，所述基于每个所述词的权重确定黑产账号聚类簇，包括：

确定所述权重高于预定权值的词所来源的所述第一字段数据文档，作为黑产数据文档；

将所述黑产数据文档对应的账号聚类簇确定为黑产账号团伙。

15、根据权利要求 9 所述的计算机设备，其中，所述基于每个所述词的权重确定黑产账号聚类簇，包括：

计算每个所述第一字段数据文档中词的权重平均值；

确定所述权重平均值高于预定平均值的所述第一字段数据文档，作为黑产数据文档；

将所述黑产数据文档对应的账号聚类簇确定为黑产账号团伙。

16、一种存储有计算机可读指令的计算机可读存储介质，所述计算机可读指令被一个或多个处理器执行时，使得一个或多个处理器执行：

确定手机号所绑定账号的数目超过预定数目时，获取所述账号的账号属性数据集，所述手机号来源于目标主体的账号数据库；

将所述账号属性数据集中的属性字段数据作为连接边，并将所述手机号作为顶点构建所述目标主体的账号检测图；

基于所述账号检测图中连接边的属性字段数据对所述账号检测图中账号进行图聚类，得到多个账号聚类簇；

利用每个所述账号聚类簇的属性字段数据，生成每个所述账号聚类簇的第一字段数据文档，并获取所述目标主体所对应白名单账号的第二字段数据文档；

根据所述第一字段数据文档及所述第二字段数据文档，计算所述第一字段数据文档中每个词的权重，所述权重指示每个所述词在所述第一字段数据文档中相对于在所述第二字段数据文档中的重要性；

基于每个所述词的权重确定黑产账号聚类簇，得到关联与所述目标主体黑产账号群体。

17、根据权利要求 16 所述的计算机可读存储介质，其中，所述确定手机号所绑定账号

的数目超过预定数目时，获取所述账号的账号属性数据集，包括：

获取目标主体与所述手机号的业务关联条件，所述业务关联条件指示所述手机号在目标业务中可以绑定账号的数目阈值，所述目标业务来源于所述目标主体；

当绑定了所述手机号的账号超过所述数目阈值时，获取所述账号的账号属性数据集。

18、根据权利要求 16 所述的计算机可读存储介质，其中，所述基于所述账号检测图中连接边的属性字段数据对所述账号检测图中账号进行图聚类，得到多个账号聚类簇，包括：

基于所述连接边的属性字段数据，对所述账号检测图利用 Connected Component 算法进行图聚类处理，得到多个账号群体；

从所述多个账号群体中，获取包含手机号个数大于等于预定个数且关联于同一登录网络地址的账号群体，得到第一账号群体组合；

从所述多个账号群体中，获取包含手机号个数大于等于预定个数且关联于同一无线网卡的物理地址的账号群体，得到第二账号群体组合；

将所述第一账号群体组合及所述第二账号群体组合确定为所述账号聚类簇。

19、根据权利要求 16 所述的计算机可读存储介质，其中，所述根据所述第一字段数据文档及所述第二字段数据文档，计算所述第一字段数据文档中每个词的权重，包括：

对于所述第一字段数据文档中的每个词，计算每个词在所述第一字段数据文档中出现的第一频率；

对于所述第一字段数据文档中的每个词，计算每个所述词在所述第一字段数据文档及所述第二字段数据文档中同时出现的第二频率；

将所述第一频率与所述第二频率的乘积，作为每个所述词的权重。

20、根据权利要求 16 所述的计算机可读存储介质，其中，所述根据所述第一字段数据文档及所述第二字段数据文档，计算所述第一字段数据文档中每个词的权重，包括：

基于公式$tf-idf(t,d)=tf(t,d)\times idf(t)$，$idf(t)=\log(N/e)$计算每个词的权重，其中，所述$tf-idf(t,d)$为权重，所述$t$为词，所述$d$为所述第一字段数据文档，所述$tf(t,d)$为所述词在所述第一字段数据文档中出现的频数，所述$idf(t)$是文档中词的逆向文本频率，所述$N$为所述第一字段数据文档及所述第二字段数据文档的总数，e 是所述第一字段数据文档及所述第二字段数据文档中出现所述词的文档数量。

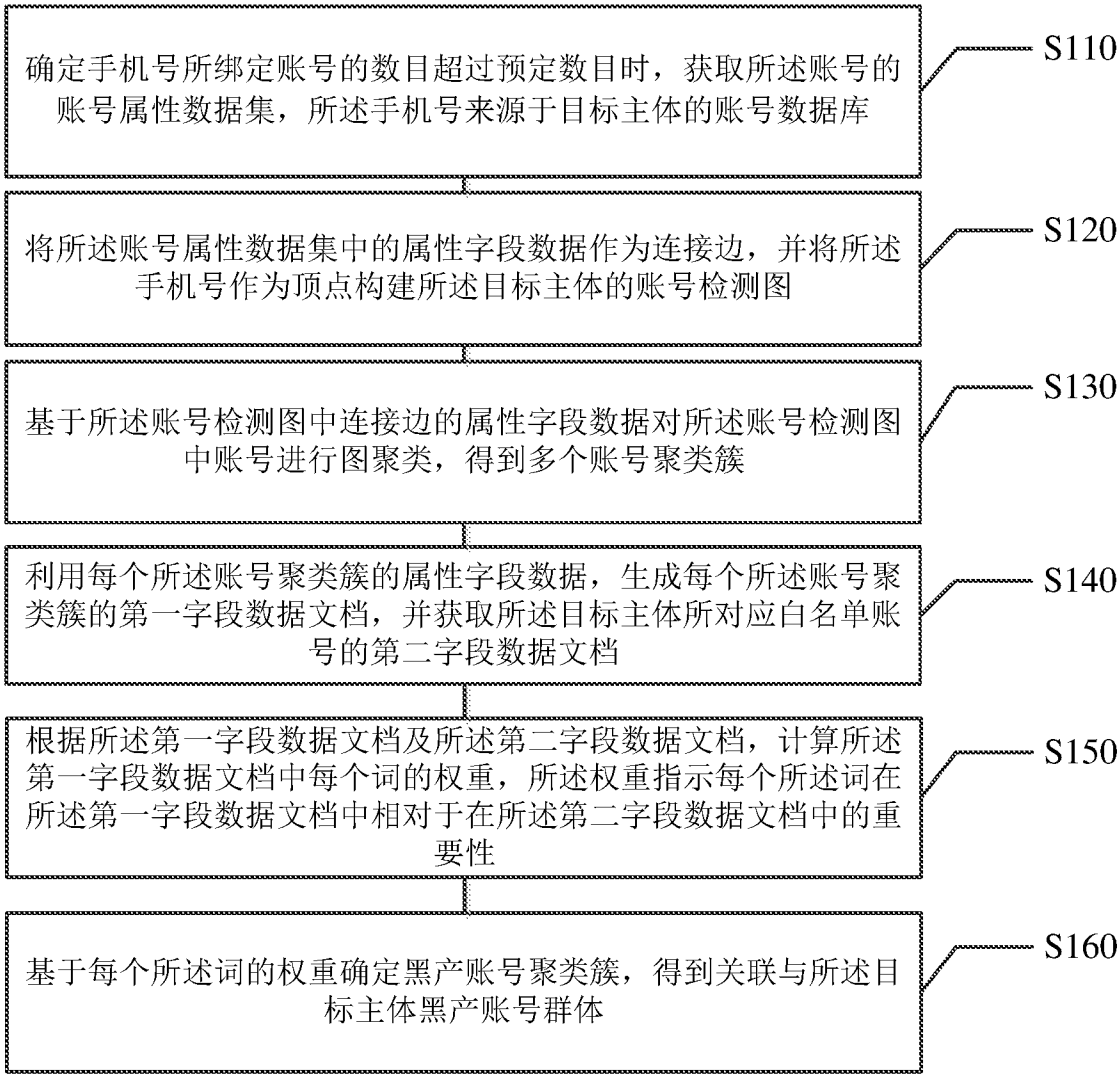
确定手机号所绑定账号的数目超过预定数目时，获取所述账号的账号属性数据集，所述手机号来源于目标主体的账号数据库
S110
将所述账号属性数据集中的属性字段数据作为连接边，并将所述手机号作为顶点构建所述目标主体的账号检测图
S120
基于所述账号检测图中连接边的属性字段数据对所述账号检测图中账号进行图聚类，得到多个账号聚类簇
S130
利用每个所述账号聚类簇的属性字段数据，生成每个所述账号聚类簇的第一字段数据文档，并获取所述目标主体所对应白名单账号的第二字段数据文档
S140
根据所述第一字段数据文档及所述第二字段数据文档，计算所述第一字段数据文档中每个词的权重，所述权重指示每个所述词在所述第一字段数据文档中相对于在所述第二字段数据文档中的重要性
S150
基于每个所述词的权重确定黑产账号聚类簇，得到关联与所述目标主体黑产账号群体
S160

图 1

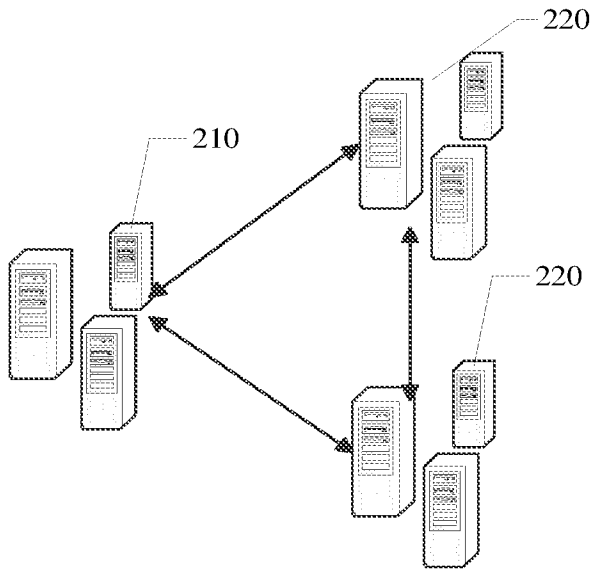
220
210
220

图 2

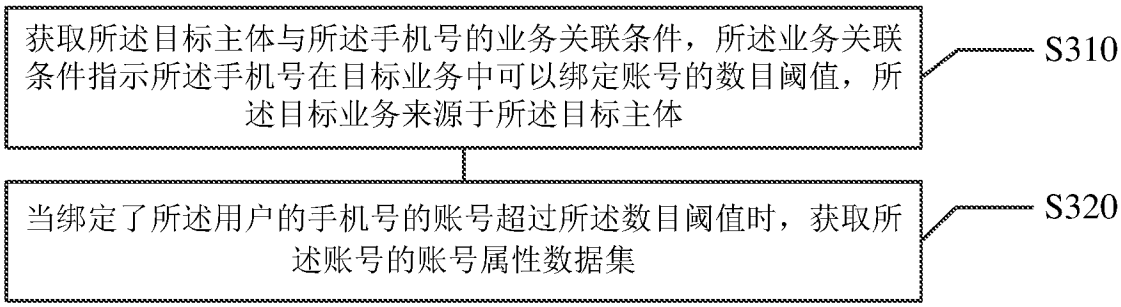
获取所述目标主体与所述手机号的业务关联条件，所述业务关联条件指示所述手机号在目标业务中可以绑定账号的数目阈值，所述目标业务来源于所述目标主体
S310
当绑定了所述用户的手机号的账号超过所述数目阈值时，获取所述账号的账号属性数据集
S320

图 3

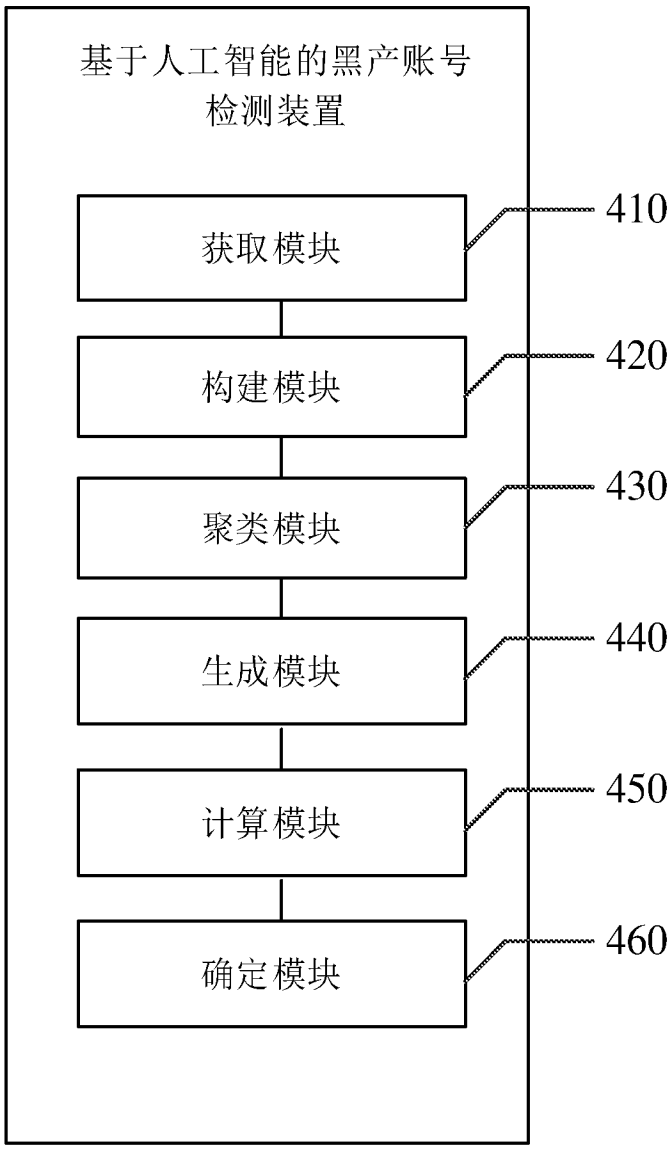
基于人工智能的黑产账号检测装置
获取模块
410
构建模块
420
聚类模块
430
生成模块
440
计算模块
450
确定模块
460

图 4

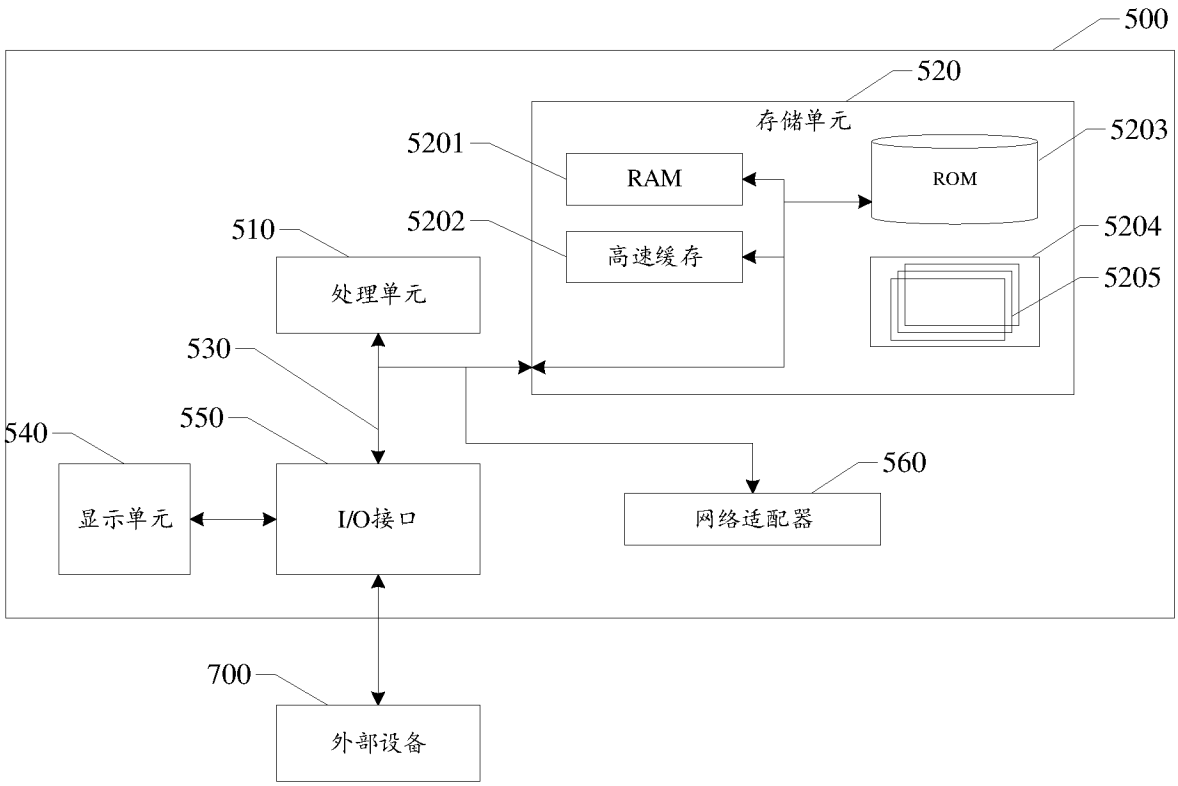
500
520
存储单元
5201
RAM
5203
ROM
5202
高速缓存
5204
5205
510
处理单元
530
550
540
560
显示单元
I/O接口
网络适配器
700
外部设备

图 5

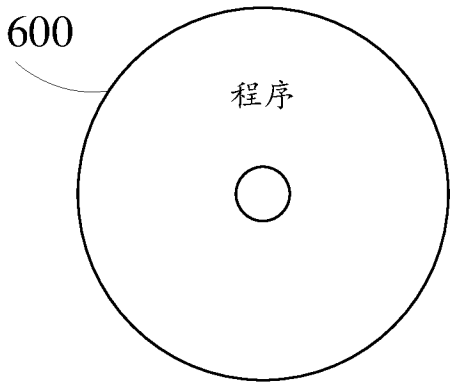
600
程序

图 6

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2021/090947

A. CLASSIFICATION OF SUBJECT MATTER

G06F 16/9535(2019.01)i

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

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06F, G06K, G06Q

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

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

CNPAT, WPI, EPODOC, CNKI: 黑产, 欺骗, 欺诈, 虚假, 异常, 恶意, 账户, 用户, 帐号, 团伙, 图, 聚类, 白名单, 重要度, 重要性, account, cheat, black production, fake, trick, graph, cluster, abnormal, white list, importance

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
PX	CN 111931048 A (PING AN TECHNOLOGY (SHENZHEN) CO., LTD.) 13 November 2020 (2020-11-13) claims 1-10	1-20
A	CN 108920947 A (BEIJING QIYI CENTURY SCIENCE & TECHNOLOGY CO., LTD.) 30 November 2018 (2018-11-30) description, paragraphs 0042-0058	1-20
A	CN 110620770 A (WEIMENG CHUANGKE NETWORK TECHNOLOGY (CHINA) CO., LTD.) 27 December 2019 (2019-12-27) description, paragraph 0030	1-20
A	CN 109660513 A (WEIMENG CHUANGKE NETWORK TECHNOLOGY (CHINA) CO., LTD.) 19 April 2019 (2019-04-19) entire document	1-20
A	CN 109948641 A (ALIBABA GROUP HOLDING LIMITED) 28 June 2019 (2019-06-28) entire document	1-20
A	US 2019318359 A1 (MASTERCARD INTERNATIONAL INCORPORATED) 17 October 2019 (2019-10-17) entire document	1-20

☐ 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
09 July 2021	**26 July 2021**

Name and mailing address of the ISA/CN	Authorized officer
China National Intellectual Property Administration (ISA/CN) **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/CN2021/090947

Patent document cited in search report	Publication date (day/month/year)	Patent family member(s)	Publication date (day/month/year)
CN 111931048 A	13 November 2020	None	
CN 108920947 A	30 November 2018	None	
CN 110620770 A	27 December 2019	None	
CN 109660513 A	19 April 2019	None	
CN 109948641 A	28 June 2019	TW 202029079 A	01 August 2020
		WO 2020147488 A1	23 July 2020
US 2019318359 A1	17 October 2019	WO 2019203947 A1	24 October 2019
		CN 110400146 A	01 November 2019

国际检索报告

国际申请号
PCT/CN2021/090947

A. 主题的分类

G06F 16/9535(2019.01)i

按照国际专利分类(IPC)或者同时按照国家分类和IPC两种分类

B. 检索领域

检索的最低限度文献(标明分类系统和分类号)

G06F，G06K，G06Q

包含在检索领域中的除最低限度文献以外的检索文献

在国际检索时查阅的电子数据库(数据库的名称，和使用的检索词（如使用））

CNPAT,WPI,EPODOC,CNKI： 黑产，欺骗，欺诈，虚假，异常，恶意，账户，用户，帐号，团伙，图，聚类，白名单，重要度，重要性，account, cheat, black production, fake, trick, graph, cluster, abnormal, white list, importance

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
PX	CN 111931048 A (平安科技深圳有限公司) 2020年 11月 13日 (2020-11-13) 权利要求1-10	1-20
A	CN 108920947 A (北京奇艺世纪科技有限公司) 2018年 11月 30日 (2018-11-30) 说明书第0042-0058段	1-20
A	CN 110620770 A (微梦创科网络科技中国有限公司) 2019年 12月 27日 (2019-12-27) 说明书第0030段	1-20
A	CN 109660513 A (微梦创科网络科技中国有限公司) 2019年 4月 19日 (2019-04-19) 全文	1-20
A	CN 109948641 A (阿里巴巴集团控股有限公司) 2019年 6月 28日 (2019-06-28) 全文	1-20
A	US 2019318359 A1 (MASTERCARD INTERNATIONAL INC.) 2019年 10月 17日 (2019-10-17) 全文	1-20

☐ 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
2021年 7月 9日	2021年 7月 26日
ISA/CN的名称和邮寄地址 中国国家知识产权局(ISA/CN) 中国北京市海淀区蓟门桥西土城路6号 100088 传真号 (86-10)62019451	受权官员 范玉霞 电话号码 86-(10)-53961331

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2021/090947

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	111931048	A	2020年 11月 13日	无			
CN	108920947	A	2018年 11月 30日	无			
CN	110620770	A	2019年 12月 27日	无			
CN	109660513	A	2019年 4月 19日	无			
CN	109948641	A	2019年 6月 28日	TW	202029079	A	2020年 8月 1日
				WO	2020147488	A1	2020年 7月 23日
US	2019318359	A1	2019年 10月 17日	WO	2019203947	A1	2019年 10月 24日
				CN	110400146	A	2019年 11月 1日