(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
2023 年 10 月 19 日 (19.10.2023)

WIPO | PCT

(10) 国际公布号
WO 2023/197350 A1

(51) 国际专利分类号:
G16H 50/80 (2018.01)

(21) 国际申请号: PCT/CN2022/088158

(22) 国际申请日: 2022 年 4 月 21 日 (21.04.2022)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
202210413620.2 2022年4月14日 (14.04.2022) CN

(71) 申请人: 浙江大学(ZHEJIANG UNIVERSITY) [CN/CN]; 中国浙江省杭州市西湖区余杭塘路866号, Zhejiang 310058 (CN)。

(72) 发明人: 贺诗波(HE, Shibo); 中国浙江省杭州市西湖区余杭塘路866号, Zhejiang 310058 (CN)。 李超(LI, Chao); 中国浙江省杭州市西湖区余杭塘路866号, Zhejiang 310058 (CN)。 余睿伟(YU, Ruiwei); 中国浙江省杭州市西湖区余杭塘路866号, Zhejiang 310058 (CN)。 吴旻诚(WU, Mincheng); 中国浙江省杭州市西湖区余杭塘路866号, Zhejiang 310058 (CN)。 陈积明(CHEN, Jiming); 中国浙江省杭州市西湖区余杭塘路866号, Zhejiang 310058 (CN)。 程鹏(CHENG, Peng); 中国浙江省杭州市西湖区余杭塘路866号, Zhejiang 310058 (CN)。 刘宇(LIU, Yu); 中国浙江省杭州市西湖区余杭塘路866号, Zhejiang 310058 (CN)。

(74) 代理人: 杭州浙科专利事务所(普通合伙) (HANGZHOU ZHEKE PATENT AGENCY (GENERAL PARTNERSHIP)); 中国浙江省杭州

WO 2023/197350 A1

(54) **Title:** WIFI MATCHING-BASED CONTACT TRACING AND PRESCREENING METHOD FOR POPULATION SUSCEPTIBLE TO INFECTIOUS DISEASE

(54) 发明名称: 一种基于WIFI匹配的传染病易感人群接触追踪预筛选方法

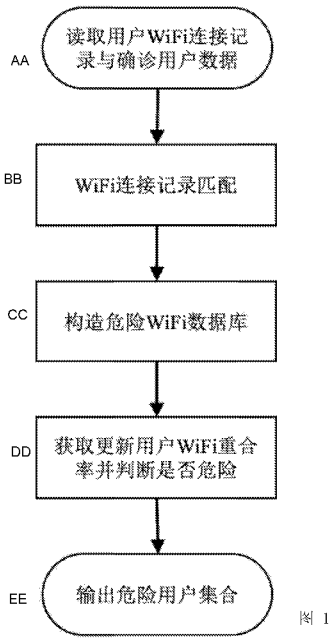


图 1

AA Read WiFi connection records of a user and data of diagnosed users
BB Match the WiFi connection records
CC Construct a risky WiFi database
DD Obtain an overlap rate of an updated user WiFi and determine whether the user is at risk
EE Output a risky user set

(57) **Abstract:** A WiFi matching-based contact tracing and prescreening method for a population susceptible to an infectious disease. In the method, on the basis of WiFi connection records of a user collected by a mobile device and a name list of anonymous identity codes of diagnosed users, the steps of record matching, information compression, constructing a risky WiFi database, calculating a risky WiFi overlap rate of the user, and the like are carried out, and finally, whether the user is at risk of infection is determined. The data required in the method is easy to obtain for a general smart mobile device, and no special application is required. Compared with a conventional

[见续页]

市拱墅区万达商业中心4幢3单元1019室 沈渊琪, Zhejiang 310015 (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, JM, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, ZA, ZM, ZW。

(84) 指定国(除另有指明，要求每一种可提供的地区保护)：ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SC, 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)。

根据细则4.17的声明：
— 发明人资格(细则4.17(iv))

本国际公布：
— 包括国际检索报告(条约第21条(3))。

contact tracing method based on GPS and Bluetooth, the method provides another dimension of information without using an additional device and sensitive data, is high in operation efficiency, and can facilitate carrying out contact tracing comprehensively and efficiently.

(57) 摘要：一种基于WiFi匹配的传染病易感人群接触追踪预筛选方法。该方法基于移动设备收集的用户WiFi连接记录与确诊用户匿名身份码名单，经过记录匹配、信息压缩、构造危险WiFi数据库、计算用户危险WiFi重合率等步骤，最终给出用户是否有感染风险的判断。其中所需的数据对于一般的智能移动设备是容易获取的，无需特殊的应用程序。相较于传统的基于GPS与蓝牙的接触追踪方法，该方法在不使用额外设备与敏感数据的基础上提供了另一个维度的信息，运行效率较高，能够帮助更全面高效地开展接触追踪。

一种基于 WiFi 匹配的传染病易感人群接触追踪预筛选方法

技术领域

本发明涉及传染病易感人群检测，尤其涉及一种基于 WiFi 匹配的传染病易感人群接触追踪预筛选方法。

背景技术

2020 年初爆发的新型冠状病毒疫情对全世界的生产生活造成了极大的冲击，截至 2022 年 3 月，全球累计确诊人数已超过 4.7 亿，累计死亡人数超过 600 万，世界卫生组织已将 2019 冠状病毒病（COVID-19）认定为史上最严重的公共卫生时间之一。新冠疫情给全球带来了巨大的生命财产损失，造成了国家、社会、人群间的严重割裂，深刻地改变着世界格局。

在防控疫情的过程中，接触追踪是十分关键的一个步骤。在公共卫生领域，接触者追踪指的是识别可能与受感染者接触的接触者的过程，以及随后收集有关这些接触者的进一步信息的过程。由于包括 COVID-19 在内的多种传染病都需要通过呼吸道传播、接触传播等方式在人群间进行传播，即发生人与人之间的接触，因此早日发现这些曾与感染者接触过的高危人群对及时隔离潜在的感染者有非常重要的作用。

传统的接触追踪方法主要依赖于对感染者的问卷调查，这种方法高度依赖于受调查者的记忆，可靠度差，同时耗费大量人工，效率低下。随着信息化时代的发展，数字接触追踪技术例如使用智能终端上的应用程序进行接触追踪等逐渐成为解决这一问题的答案。数字接触追踪通常需要借助位置数据，如 GPS、WiFi、通信基站、蓝牙信标等，根据使用数据类型又可进一步分为两类方法：

1) 使用绝对位置数据的方法：使用绝对的位置信息，如 GPS 位置、相对于静态的 WiFi 热点或通信基站的位置等数据，这些数据通常可以表示为地理坐标的形式。这些数据对于传染病接触追踪可能精确率不够高，但是通过大量此类数据能够发现特定用户的行动模式从而进行聚类。

2) 使用相对位置数据的方法：使用相对的位置信息，如两台开启了蓝牙的设备的交互数据、同时使用同一台公共交通工具如飞机、火车等的数据，这些数据通常表示为在某一时间点上位于同一区域的数据对。这些数据相对而言有更高的精确率，但一方面需要特定设备（如蓝牙），另一方面过于关注点与点的互动，而缺乏整体的信息。

两类数据都有了一些实际应用。如中国普遍使用的健康码就利用了用户的 GPS 序列数据进行接触判断与追踪，而新加坡政府提供的 TraceTogether 服务就通过强制要求公民下载专门的使用蓝牙信号的应用程序或佩戴专门的蓝牙通信设备来进行接触追踪，但两种方案在应用上都有其缺陷：GPS 数据在室内精度有限，且涉及到人的具体位置轨迹等敏感信息；蓝牙点对点接触数据需要大规模使用专业设备，成本高昂不易普及。

针对上述方法的缺点，申请人提出一种基于 WiFi 匹配的传染病易感人群接触追踪预筛选方法。

发明内容

本发明目的在于对现有应用技术的缺陷加以弥补，通过使用不同类型的数据，提供更全面的传染病易感人群接触追踪方法，该方法可以利用历史传染病确诊患者的 WiFi 接入记录构造危险 WiFi 数据库，在获取每日更新的用户 WiFi 接入记录后给出用户的感染风险，提高寻找易感人群的效率。

本发明的目的通过以下的技术方案实现：

一种基于 WiFi 匹配的传染病易感人群接触追踪预筛选方法，包括以下步骤：

1) 用户匿名身份码记录匹配：通过数据库获取确诊用户唯一匿名身份码数据表，利用 hive sql 搜索这些 id 在总数据库中对应的 WiFi 连接记录，按照日期建立数据表分区，保存为原始数据记录，数据时间跨度总计T天，构造确诊用户 WiFi 连接数据集$R = \{R_1, R_2, ..., R_T\}$，其中第$i$天的记录为$R_i$，总计有$k_i$条记录，则$R_i = \{r_{i1}, r_{i2}, \ldots, r_{ik_i}\}$，其中每条记录$r_{ij} = (gid_{ij}, confirmDay_{ij}, mac_{ij})$，$gid_{ij}$表示用户唯一匿名身份码、$confirmDay_{ij}$表示用户确诊日期、$mac_{ij}$表示用户连接 WiFi 的 mac 地址；

2) 数据预处理：对步骤(1)中获取的数据集R保存至代码本地并存储，为节省存储空间，建立双向映射字典类，在获取某字符串后，根据内部已存储条目数量针对该字符串建立新的指代符号，将该字符串与指代符号分别作为键与值建立映射；建立用户匿名身份码的代码双向映射字典$dict_{user}$，记用户总数为N_u，将原本的用户统一长度的匿名身份码映射为代号u_{i_u}，其中$i_u = 1,2,3, ..., N_u$；建立 WiFi 的代码双向映射字典$dict_{wifi}$，记 WiFi 总数为N_w，将原本的 WiFi 的 mac 地址映射为代号w_{u_w}，其中$u_w = 1,2,3, ..., N_w$；原始记录数据集映射后得到的记录数据集记为$\overline{R}$；

3) 构造危险 WiFi 数据库：利用步骤(2)中经过代号映射后的信息数据集$\overline{R}$，记危险 WiFi

数据库为$dict_{dan}$，提取$\overline{R}$中确诊用户的 WiFi 连接记录，取记录中的 WiFi 代号w_{u_w}放入$dict_{dan}$，同时记录该 WiFi 收入危险 WiFi 数据库中的日期为当前日期day_{now}；若该 WiFi 已经存在，则改为更新其收入日期为day_{now}，然后将收入日期day_{u_w}与当前日期day_{now}之差超过危险 WiFi 消失阈值thr_{dan}的记录删去；

4) 感染风险判断：记每日新获取的用户 WiFi 连接记录$\widetilde{R}$，将$\widetilde{R}$按用户进行分组，对每一用户，将其当日连接的所有 WiFi 与$dict_{dan}$进行匹配，记重合率$r_{user} = |C_{user}|/|W_{user}|$，其中$|C_{user}|$为该用户当日连接的在$dict_{dan}$中的 WiFi 数量，$|W_{user}|$为该用户当日连接的 WiFi 总数，若$r_{user}$高于判断阈值$thr_{user}$，则判断此用户为危险用户。

进一步地，所述步骤 1)用户匿名身份码记录匹配中，使用 hive sql 工具将确诊用户唯一匿名身份码与全国用户 WiFi 连接记录总数据库进行匹配，并生成用于后续处理分区确诊用户数据记录表，具体实施方法如下：

1.1) 原始数据库包含两部分：1.1.1)确诊用户信息数据表，包括确诊用户匿名身份码 gid_confirm、确诊日期 confirm_day；1.1.2)全国用户 WiFi 连接记录数据表，包括用户匿名身份码 gid、连接 WiFi 的 mac 地址 mac、记录生成日期 day、记录生成地区 area、记录生成时刻 hour、记录时刻内 WiFi 连接次数 times，以两个数据库的相同的用户匿名身份码为索引，利用 hive sql 工具生成确诊用户 WiFi 连接记录数据表；

其中全国用户 WiFi 连接记录数据表按地区 area 与日期 day 进行分区，在构造目标数据表时，为提高筛选效率、减少筛选时间，需要提前确定目标筛选地区（其地区代码为 AREACORE）与目标考察的起止时间（分别为 BEGINDAY 与 ENDDAY）；

1.2) 为防止后续处理确诊用户 WIFI 连接记录数据表因数据量大导致内存溢出，使用 hive sql 工具将数据表依照日期为索引进行分区处理。

进一步地，所述步骤 2)数据预处理中，使用双向映射字典对原始用户匿名身份码与 WiFi 的 mac 地址进行压缩与还原，具体实施方法如下：

2.1) 构造双向映射字典类 TwoWayDict，其特点在于输入后原始未压缩字符串，字典内部形成未压缩字符串到压缩字符串、压缩字符串到未压缩字符串的双向映射，即将未压缩字符串与压缩字符串各作为键与对应的值保存，以在后续处理中节省存储空间与内存，当获取预测结果后，同样根据该双向字典获取原始的未压缩字符串；

2.2) 分别将用户匿名身份码与 WiFi 的 mac 地址以上述方法建立两个双向映射字典$dict_{user}$与$dict_{wifi}$。

进一步地，所述步骤 3)构造危险 WiFi 数据库中，依照日期将历史用户 WiFi 连接数据中

的 mac 地址加入危险 WiFi 数据库，并动态删除过期 WiFi 以降低假阳性率，具体实施方法如下：

3.1) 处理历史数据，提取历史确诊用户 WiFi 连接记录中的 WiFi 的 mac 地址与确诊日期，将其放入危险 WiFi 数据库$dict_{dan}$中，$dict_{dan}$以 WiFi 的 mac 地址为键，以确诊日期为值，有重复 mac 地址的，选择日期最近的保存为值；

3.2) 处理每日更新数据，当获取每日更新数据后，将当日确诊患者所连接的 WiFi 的 mac 地址以上述方法加入$dict_{dan}$中，添加完毕后遍历$dict_{dan}$，如果有 mac 地址对应的确诊日期与当前日期的差值超过危险 WiFi 消失阈值thr_{dan}，则将其从危险 WiFi 数据库中删去。

进一步地，所述步骤 4)感染风险判断中，将每日更新的用户 WiFi 连接数据与危险 WiFi 数据库进行比对获取重合率，并根据阈值判断是否为危险用户，具体步骤如下：

4.1) 获取每日更新的用户 WiFi 连接记录后，按照用户匿名身份码进行分组，计算其当日连接 WiFi 与危险 WiFi 数据库的重合率，用户 user 的重合率r_{user}计算公式为：

$$C_{user} = W_{user} \cap W_{dan}$$

$$r_{user} = \frac{|C_{user}|}{|W_{user}|}$$

其中W_{user}为该用户当日连接的 WiFi，W_{dan}为危险 WiFi 数据库，$|S|$表示集合S的元素个数；

若R_{user}高于危险判断阈值thr_{user}，将其判定为危险用户，应对其进行重点筛查。

本发明一种基于 WiFi 匹配的传染病易感人群接触追踪预筛选方法，主要通基于移动设备收集的用户 WiFi 连接记录与确诊用户匿名身份码名单，经过记录匹配、信息压缩、构造危险 WiFi 数据库、计算用户危险 WiFi 重合率等步骤得到最终的危险用户名单。本发明对 WiFi 连接记录数据进行分析、修正、降维，相比于接触追踪方法而言可以在不涉及位置等隐私信息，不使用额外设备的同时节省计算资源，减少处理时间，提高筛查效率，为后续寻找可能与确诊用户发生接触的高危易感人群提供了另一个维度更全面的数据支撑。

附图说明

图1是本发明的基于WiFi匹配的数字接触追踪预筛选的方法流程图；

图2是本发明在实施例用户群中筛选的精确率与召回率比较结果图。

具体实施方式

以下结合附图对本发明的具体实施方法与工作原理作如下详述：

实施例

本实施例对从某地获取的自2020年1月1日至2020年3月20日采集到的用户WiFi连接记录数据以及该时间段内的某呼吸道传染病确诊用户匿名信息。数据集包括的具体变量以及相关数据信息如表1、表2、表3、表4所示：

表1 某地部分用户WiFi连接记录数据

gid	hour	mac	times	day	area
ANDROID-bd78278072f845619d3653d4cf78ca51	11	9ca61526d6c7	1	20200202	4***0000
ANDROID-b57e2d5b37967b2fa289cbeef989f8e4	10	1c60de4cee34	1	20200202	4***0000
ANDROID-89a360d236df804acf7446fa2843335c	22	0cd86cde470a	1	20200202	4***0000
ANDROID-4d341ee77d756b29b4bfc58347707cfd	12	8cab8ec48890	1	20200202	4***0000
ANDROID-c0318a70d23f46f890d38aa92a3853fb	03	fcd733b93950	1	20200202	4***0000
…	…	…	…	…	…

表2 某地部分用户WiFi连接记录数据字段说明

字段名	数据类型	说明
gid	string	用户匿名身份码
hour	int	生成记录的时间段
mac	string	连接WiFi的mac地址
times	int	记录生成时间段内连接该WiFi的次数
day	int	记录生成的日期
area	int	记录所属地区的区域代码

表3 全国确诊用户数据

confirm_gid	confirm_day	day
ANDROID-863cc84e5e0548b696eb0c426f249975	20200315	20200315
ANDROID-7af1888075bf4e439a0fefd4b52a865a	20200315	20200315
IOS-5eb8760cc460319162e31f888bfae1ff	20200315	20200315
ANDROID-8d2c876ddf894639a1a80f3ebb695569	20200315	20200315
ANDROID-5c4f0b105585408b9186dba9454ddd58	20200315	20200315

表4 移动设备数据集中wifi_list变量字段解释

字段名	数据类型	说明
confirm_gid	string	确诊用户匿名身份码
confirm_day	int	该用户确诊日期
day	int	该记录所属日期

本实施例中默认传染病易感人群接触追踪预筛选方法的实施数据集即为上述某地用户的WiFi 连接数据与某传染病确诊用户数据，方法结果为得到的危险用户列表，其详细实施步骤具体如下：

1) 根据用户匿名身份码匹配 WiFi 连接记录，利用 hive sql 工具，根据记录日期与所选地区代码从全国用户 WiFi 连接记录总表中提取子表以提高匹配效率，以用户匿名身份码相同为筛选条件，保留对应记录中的用户匿名身份码、连接 WiFi 的 mac 地址、记录生成日期、用户确诊日期字段，并以记录生成日期为分区索引，对生成的确诊用户 WiFi 连接记录数据表进行分区处理。为了验证预测效果，本实施例中采样了部分健康用户的记录进行同样的处理，健康用户的确诊日期字段做缺省处理。同时在表中添加 seed 字段，seed 等于 0 表示健康用户，seed 等于 1 表示确诊用户。

2) 利用步骤1)中获取的确诊用户WiFi连接记录数据表按照日期依序处理。当读入数据后，分别提取其gid列与mac地址列，分别构造双向映射字典$dict_{user}$与$dict_{wifi}$，对gid字符串与mac地址字符串进行压缩，压缩后的数据表如表5所示：

表5 经双向映射字典压缩后数据表

gid	mac	times	quezhen_date	seed
u1	w93	13	\N	0
u1	w865054	11	\N	0
u1	w2255432	15	\N	0
u1	w96	9	\N	0
u1	w865074	6	\N	0

3) 处理历史数据，考察所有确诊用户连接的 WiFi，考虑某传染病存在潜伏期，即感染后表现为无症状直至确诊，判定记录产生日期若在确诊日期T_{inc}内则将该 WiFi 的 mac 地址与当天日期加入危险 WiFi 数据库中，本实施例中取T_{inc}为 7 天。获得的危险 WiFi 数据库形式如下：

{'w3706696': datetime.date(2020, 3, 18),
'w3658960': datetime.date(2020, 2, 16),
'w3737021': datetime.date(2020, 3, 11),
'w3834102': datetime.date(2020, 3, 15),
'w194670': datetime.date(2020, 3, 20),
'w2787916': datetime.date(2020, 2, 23),…}

4) 判断每日更新数据，将每日更新的用户 WiFi 连接记录按用户 gid 进行分组，并汇总每一用户当日扫描的 WiFi，计算与危险 WiFi 的重合率，根据重合率是否高于阈值thr_{user}给出是否为危险用户的判断，本实施例中thr_{user}设为 0.1，构造当日的危险用户集合，其形式如下：

{'u46515', 'u19505', 'u12248', 'u35433', 'u19808', …}

5) 根据当日实际确诊的用户信息更新危险 WiFi 数据库，已经在危险 WiFi 数据库内的则更新其记录日期。

6) 将当前的压缩用户 id 映射回原始用户匿名身份码，对这些用户进行重点筛查。如步骤 4)中用户代码映射结果为：

ANDROID-21010e17edd0437599c2388cd65d130b
ANDROID-c8607f33b36a40b7b17c227867fb290e
ANDROID-7e2f8561d05f468ca553ae9daa55b18c
ANDROID-018570e1daa544f082b655cf80feae54
ANDROID-cb2c78a391ac47229f2aa499d20a2a5f

(7) 本实施例中，每日精确率与召回率如图 1 所示。考虑到传染病筛查相比于精确率，召回率指标更为重要，因此本发明较好地实现了危险人员的筛查。

本发明一种基于 WiFi 匹配的传染病易感人群接触追踪预筛选方法，主要通基于移动设备收集的用户 WiFi 连接记录与确诊用户匿名身份码名单，经过记录匹配、信息压缩、构造危险 WiFi 数据库、计算用户危险 WiFi 重合率等步骤得到最终的危险用户名单。图 1 为基于 WiFi

匹配的传染病易感人群接触追踪预筛选方法的具体流程。整个实施例按照图 1 中所示的流程，对用户 WiFi 连接记录数据集进行处理并最终得出危险人员筛查结果。图 2 为本方法在用户群中筛选的精确率与召回率比较结果，对 WiFi 连接记录数据进行分析、修正、降维相比于接触追踪方法而言可以在不涉及位置等隐私信息、不使用额外设备的同时节省计算资源，减少处理时间，提高筛查效率，为后续寻找可能与确诊用户发生接触的高危易感人群提供了另一个维度更全面的数据支撑。

上述实施例只是本发明的举例，尽管为说明目的公开了本发明的最佳实例和附图，但是本领域的技术人员可以理解：在不脱离本发明及所附的权利要求的精神和范围内，各种替换、变化和修改都是可能的。因此，本发明不应局限于最佳实施例和附图所公开的内容。

权 利 要 求 书

1.一种基于 WiFi 匹配的传染病易感人群接触追踪预筛选方法，其特征在于，包括以下步骤：

1) 用户匿名身份码记录匹配：通过数据库获取确诊用户唯一匿名身份码数据表，利用 hive sql 搜索这些 id 在总数据库中对应的 WiFi 连接记录，按照日期建立数据表分区，保存为原始数据记录，数据时间跨度总计T天，构造确诊用户 WiFi 连接数据集$R = \{R_1, R_2, ..., R_T\}$，其中第$i$天的记录为$R_i$，总计有$k_i$条记录，则$R_i = \{r_{i1}, r_{i2}, \ldots, r_{ik_i}\}$，其中每条记录$r_{ij} = (gid_{ij}, confirmDay_{ij}, mac_{ij})$，$gid_{ij}$表示用户唯一匿名身份码、$confirmDay_{ij}$表示用户确诊日期、$mac_{ij}$表示用户连接 WiFi 的 mac 地址；

2) 数据预处理：对步骤(1)中获取的数据集R保存至代码本地并存储，建立双向映射字典类，在获取任一字符串后，根据内部已存储条目数量针对该字符串建立新的指代符号，将该字符串与指代符号分别作为键与值建立映射；建立用户匿名身份码的代码双向映射字典$dict_{user}$，记用户总数为N_u，将原本的用户统一长度的匿名身份码映射为代号u_{i_u}，其中$i_u = 1,2,3, ..., N_u$；建立 WiFi 的代码双向映射字典$dict_{wifi}$，记 WiFi 总数为N_w，将原本的 WiFi 的 mac 地址映射为代号w_{u_w}，其中$u_w = 1,2,3, ..., N_w$；原始记录数据集映射后得到的记录数据集记为$\overline{R}$；

3) 构造危险 WiFi 数据库：利用步骤(2)中经过代号映射后的信息数据集$\overline{R}$，记危险 WiFi 数据库为$dict_{dan}$，提取$\overline{R}$中确诊用户的 WiFi 连接记录，取记录中的 WiFi 代号w_{u_w}放入$dict_{dan}$，同时记录该 WiFi 收入危险 WiFi 数据库中的日期为当前日期day_{now}；若该 WiFi 已经存在，则改为更新其收入日期为day_{now}，然后将收入日期day_{u_w}与当前日期day_{now}之差超过危险 WiFi 消失阈值thr_{dan}的记录删去；

4) 感染风险判断：记每日新获取的用户 WiFi 连接记录$\widetilde{R}$，将$\widetilde{R}$按用户进行分组，对每一用户，将其当日连接的所有 WiFi 与$dict_{dan}$进行匹配，记重合率$r_{user} = |C_{user}|/|W_{user}|$，其中$|C_{user}|$为该用户当日连接的在$dict_{dan}$中的 WiFi 数量，$|W_{user}|$为该用户当日连接的 WiFi 总数，若$r_{user}$高于判断阈值$thr_{user}$，则判断此用户为危险用户。

2.根据权利要求 1 中所述的一种基于 WiFi 匹配的传染病易感人群接触追踪预筛选方法，其特征在于所述步骤 1)用户匿名身份码记录匹配中，使用 hive sql 工具将确诊用户唯一匿名

身份码与全国用户 WiFi 连接记录总数据库进行匹配，并生成用于后续处理分区确诊用户数据记录表，具体实施方法如下：

1.1) 原始数据库包含两部分：1.1.1)确诊用户信息数据表，包括确诊用户匿名身份码 gid_confirm、确诊日期 confirm_day；1.1.2)全国用户 WiFi 连接记录数据表，包括用户匿名身份码 gid、连接 WiFi 的 mac 地址 mac、记录生成日期 day、记录生成地区 area、记录生成时刻 hour、记录时刻内 WiFi 连接次数 times，以两个数据库的相同的用户匿名身份码为索引，利用 hive sql 工具生成确诊用户 WiFi 连接记录数据表；

其中全国用户 WiFi 连接记录数据表按地区 area 与日期 day 进行分区，在构造目标数据表时提前确定目标筛选地区与目标考察的起止时间；

1.2) 使用 hive sql 工具将数据表依照日期为索引进行分区处理。

3.根据权利要求 1 中所述的一种基于 WiFi 匹配的传染病易感人群接触追踪预筛选方法，其特征在于所述步骤 2)数据预处理中，使用双向映射字典对原始用户匿名身份码与 WiFi 的 mac 地址进行压缩与还原，具体实施方法如下：

2.1) 构造双向映射字典类 TwoWayDict，其特点在于输入后原始未压缩字符串，字典内部形成未压缩字符串到压缩字符串、压缩字符串到未压缩字符串的双向映射，即将未压缩字符串与压缩字符串各作为键与对应的值保存，以在后续处理中节省存储空间与内存，当获取预测结果后，同样根据该双向字典获取原始的未压缩字符串；

2.2) 分别将用户匿名身份码与 WiFi 的 mac 地址建立两个双向映射字典$dict_{user}$与$dict_{wifi}$。

4.根据权利要求 1 中所述的一种基于 WiFi 匹配的传染病易感人群接触追踪预筛选方法，其特征在于所述步骤 3)构造危险 WiFi 数据库中，依照日期将历史用户 WiFi 连接数据中的 mac 地址加入危险 WiFi 数据库，并动态删除过期 WiFi 以降低假阳性率，具体实施方法如下：

3.1) 处理历史数据，提取历史确诊用户 WiFi 连接记录中的 WiFi 的 mac 地址与确诊日期，将其放入危险 WiFi 数据库$dict_{dan}$中，$dict_{dan}$以 WiFi 的 mac 地址为键，以确诊日期为值，有重复 mac 地址的，选择日期最近的保存为值；

3.2) 处理每日更新数据，当获取每日更新数据后，将当日确诊患者所连接的 WiFi 的 mac 地址以上述方法加入$dict_{dan}$中，添加完毕后遍历$dict_{dan}$，如果有 mac 地址对应的确诊日期与当前日期的差值超过危险 WiFi 消失阈值thr_{dan}，则将其从危险 WiFi 数据库中删去。

5.根据权利要求 1 中所述的一种基于 WiFi 匹配的传染病易感人群接触追踪预筛选方法，其特征在于所述步骤 4)感染风险判断中，将每日更新的用户 WiFi 连接数据与危险 WiFi 数据库进行比对获取重合率，并根据阈值判断是否为危险用户，具体步骤如下：

4.1) 获取每日更新的用户 WiFi 连接记录后，按照用户匿名身份码进行分组，计算其当日连接 WiFi 与危险 WiFi 数据库的重合率，用户 user 的重合率r_{user}计算公式为：

$$C_{user} = W_{user} \cap W_{dan}$$

$$r_{user} = \frac{|C_{user}|}{|W_{user}|}$$

其中W_{user}为该用户当日连接的 WiFi，W_{dan}为危险 WiFi 数据库，$|S|$表示集合S的元素个数；

若R_{user}高于危险判断阈值thr_{user}，将其判定为危险用户，应对其进行重点筛查。

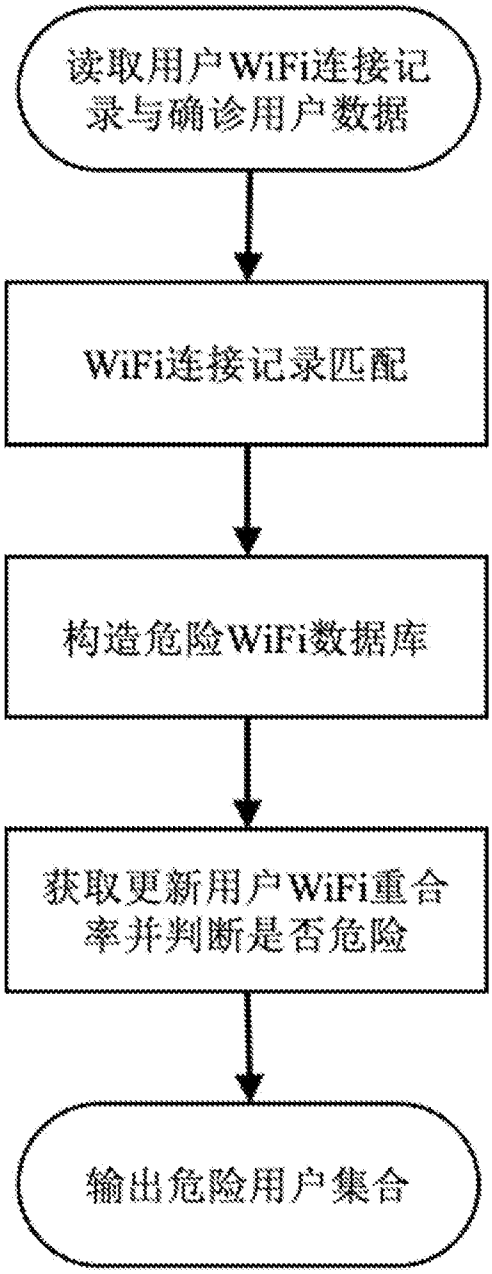

读取用户WiFi连接记录与确诊用户数据
WiFi连接记录匹配
构造危险WiFi数据库
获取更新用户WiFi重合率并判断是否危险
输出危险用户集合

图 1

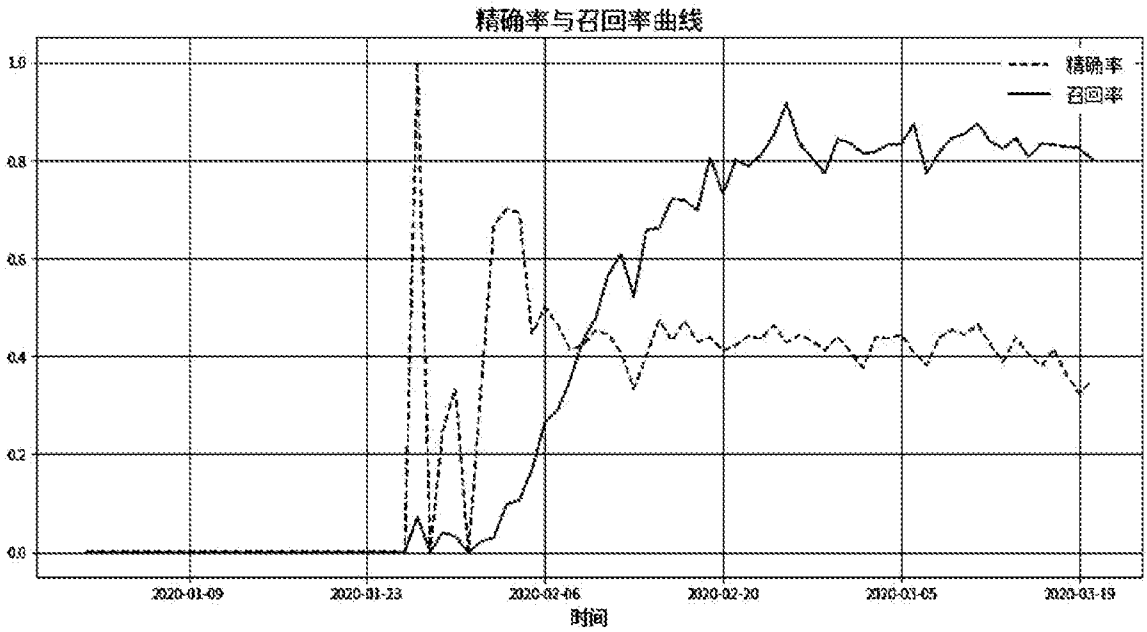

精确率与召回率曲线
精确率
召回率
1.0
0.8
0.6
0.4
0.2
0.0
2020-01-09
2020-01-23
2020-02-06
2020-02-20
2020-03-05
2020-03-19
时间

图 2

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2022/088158

A. CLASSIFICATION OF SUBJECT MATTER

G16H 50/80(2018.01)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)

G16H

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, CNKI, WPI, EPODOC: 无线网络, 传染病, 流行病, 流感, 筛选, 接触, 密接, 追踪, 预测, 感染, 确诊, wireless network, wifi, infectious disease, epidemic, flu, influenza, screen+, contact+, track+, forecast+, infect+, diagnosis

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	CN 111354472 A (DAI JIANRONG et al.) 30 June 2020 (2020-06-30) description, paragraphs [0034]-[0060] and [0153], and figures 1-13	1-5
A	CN 111354473 A (TIANJIN LVZHOU ENERGY EQUIPMENT CO., LTD.) 30 June 2020 (2020-06-30) entire document	1-5
A	CN 111446006 A (CHEN TIANHUI) 24 July 2020 (2020-07-24) entire document	1-5
A	CN 104778642 A (BEIJING UNIVERSITY OF POSTS AND TELECOMMUNICATIONS) 15 July 2015 (2015-07-15) entire document	1-5
A	US 2019252078 A1 (X DEVELOPMENT LLC) 15 August 2019 (2019-08-15) entire document	1-5
A	WO 2021167159 A1 (TINA3D CO., LTD.) 26 August 2021 (2021-08-26) entire document	1-5

☐ 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
20 December 2022	**29 December 2022**
Name and mailing address of the ISA/CN **China National Intellectual Property Administration (ISA/CN)** **No. 6, Xitucheng Road, Jimenqiao, Haidian District, Beijing 100088, China** Facsimile No. **(86-10)62019451**	Authorized officer Telephone No.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2022/088158

Patent document cited in search report	Publication date (day/month/year)	Patent family member(s)	Publication date (day/month/year)
CN 111354472 A	30 June 2020	None	
CN 111354473 A	30 June 2020	None	
CN 111446006 A	24 July 2020	None	
CN 104778642 A	15 July 2015	CN 104778642 B	13 April 2018
US 2019252078 A1	15 August 2019	None	
WO 2021167159 A1	26 August 2021	KR 102159722 B1	24 September 2020

国际检索报告

国际申请号
PCT/CN2022/088158

A. 主题的分类

G16H 50/80(2018.01)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)

G16H

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, CNKI, WPI, EPODOC：无线网络，传染病，流行病，流感，筛选，接触，密接，追踪，预测，感染，确诊，wireless network, wifi, infectious disease, epidemic, flu, influenza, screen+, contact+, track+, forecast+, infect+, diagnosis

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
A	CN 111354472 A (戴建荣 等) 2020年6月30日 (2020-06-30) 说明书第[0034]-[0060]，[0153]段，图1-13	1-5
A	CN 111354473 A (天津绿州能源装备有限公司) 2020年6月30日 (2020-06-30) 全文	1-5
A	CN 111446006 A (陈恬慧) 2020年7月24日 (2020-07-24) 全文	1-5
A	CN 104778642 A (北京邮电大学) 2015年7月15日 (2015-07-15) 全文	1-5
A	US 2019252078 A1 (X DEVELOPMENT L.L.C.) 2019年8月15日 (2019-08-15) 全文	1-5
A	WO 2021167159 A1 (TINA3D CO., LTD.) 2021年8月26日 (2021-08-26) 全文	1-5

☐ 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
2022年12月20日	2022年12月29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 张玥 电话号码 86-(10)-53962603

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2022/088158

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 111354472 A	2020年6月30日	无	
CN 111354473 A	2020年6月30日	无	
CN 111446006 A	2020年7月24日	无	
CN 104778642 A	2015年7月15日	CN 104778642 B	2018年4月13日
US 2019252078 A1	2019年8月15日	无	
WO 2021167159 A1	2021年8月26日	KR 102159722 B1	2020年9月24日