

(12) 按照专利合作条约所公布的国际申请

(19) 世界知识产权组织
国际局

(43) 国际公布日
2019 年 4 月 18 日 (18.04.2019)



(10) 国际公布号
WO 2019/071897 A1

(51) 国际专利分类号:
G06F 17/30 (2006.01)

(21) 国际申请号: PCT/CN2018/076173

(22) 国际申请日: 2018 年 2 月 10 日 (10.02.2018)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
201710953455.9 2017 年 10 月 13 日 (13.10.2017) 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八卦岭工业区平安大厦六楼, Guangdong 518000 (CN)。

(72) 发明人: 许开河(XU, Kaihe); 中国广东省深圳市福田区八卦岭工业区平安大厦六楼, Guangdong

518000 (CN)。 兰相如(LAN, Xiangru); 中国广东省深圳市福田区八卦岭工业区平安大厦六楼, Guangdong 518000 (CN)。 王建明(WANG, Jianming); 中国广东省深圳市福田区八卦岭工业区平安大厦六楼, Guangdong 518000 (CN)。 肖京(XIAO, Jing); 中国广东省深圳市福田区八卦岭工业区平安大厦六楼, Guangdong 518000 (CN)。

(74) 代理人: 深圳市沃德知识产权代理事务所(普通合伙)(SHENZHEN WORLD INTELLECTUAL PROPERTY AGENCY (GENERAL PARTNERSHIP)); 中国广东省深圳市福田区园岭街道八卦四路 10 号中浩大厦 1528-1530 室于志光, 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,

(54) Title: REAL-TIME RECOMMENDATION METHOD, ELECTRONIC DEVICE, AND COMPUTER READABLE STORAGE MEDIUM

(54) 发明名称: 实时推荐方法、电子设备及计算机可读存储介质

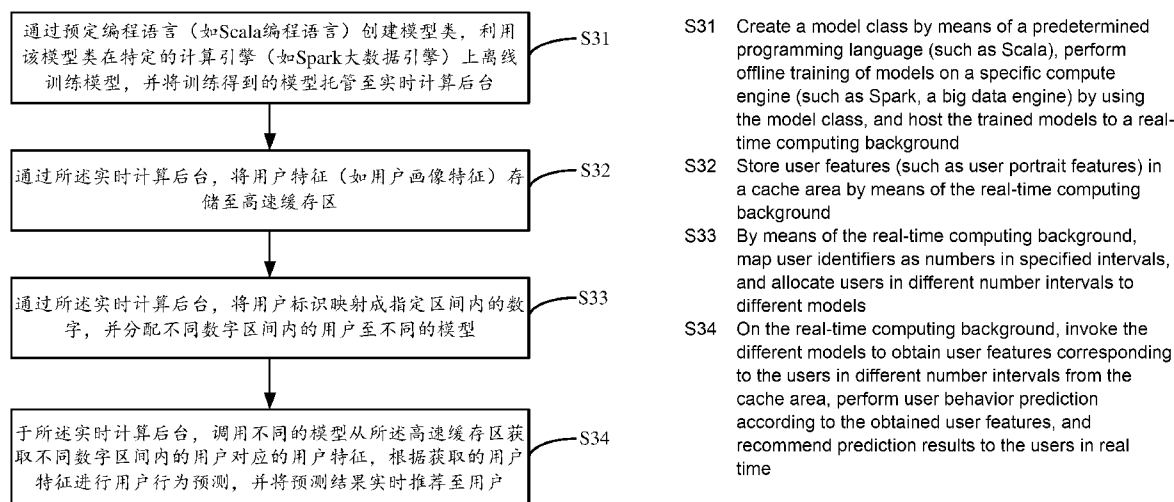


图 3

(57) Abstract: A real-time recommendation method. The method comprises the steps of: creating a model class by means of a predetermined programming language, performing offline training of models on a specific compute engine by using the model class, and hosting the trained models to a real-time computing background (S31); storing user features in a cache area by means of the real-time computing background (S32); mapping user identifiers as numbers in specified intervals, and allocating users in different number intervals to different models (S33); and invoking the different models to obtain user features corresponding to the users in different number intervals from the cache area, performing user behavior prediction according to the obtained user features, and recommending prediction results to the users in real time (S34). An iterative model can be quickly updated, so that scale-out is facilitated.



CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW。

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

本国际公布:

— 包括国际检索报告(条约第21条(3))。

(57) 摘要: 一种实时推荐方法, 该方法包括步骤: 通过预定编程语言创建模型类, 利用该模型类在特定的计算引擎上离线训练模型, 并将训练得到的模型托管至实时计算后台(S31); 通过所述实时计算后台, 将用户特征存储至高速缓存区(S32); 将用户标识映射成指定区间内的数字, 并分配不同数字区间内的用户至不同的模型(S33); 调用不同的模型从所述高速缓存区获取不同数字区间内的用户对应的用户特征, 根据获取的用户特征进行用户行为预测, 并将预测结果实时推荐至用户(S34)。可以快速更新迭代模型, 方便横向扩展。

实时推荐方法、电子设备及计算机可读存储介质

本申请要求于2017年10月13日提交中国专利局、申请号为201710953455.9、发明名称为“实时推荐方法、电子设备及计算机可读存储介
5 质”的中国专利申请的优先权，其全部内容通过引用结合在申请中。

技术领域

本申请涉及计算机信息技术领域，尤其涉及一种实时推荐方法、电子设备及计算机可读存储介质。

10

背景技术

目前，多数大数据引擎（如Spark引擎）的数据分析工具只能处理离线的数据，对应的实时处理组件（如spark-streaming组件）需要大数据引擎的集群支持，导致成本高、耗时长（100ms-2s），不能满足实时推荐低耗时的要求。

15 另外，目前的大数据引擎训练出来的推荐模型不能及时更新到线上产生效益，并且多个模型同时在线上进行A/B测试也不方便进行管理。故，现有技术中的实时推荐方法设计不够合理，亟需改进。

发明内容

20 有鉴于此，本申请提出一种实时推荐方法、电子设备及计算机可读存储介质，通过实时计算后台的运作缩短了预测时间，并利用前端界面系统实现了同时对多个模型进行线上管理。

首先，为实现上述目的，本申请提出一种电子设备，所述电子设备包括存储器和处理器，所述存储器上存储有可在所述处理器上运行的实时推荐系
25 统，所述实时推荐系统被所述处理器执行时实现如下步骤：

通过预定编程语言创建模型类，利用该模型类在特定的计算引擎上离线

训练模型，并将训练得到的模型托管至实时计算后台；

通过所述实时计算后台，将用户特征存储至高速缓存区；

通过所述实时计算后台，将用户标识映射成指定区间内的数字，并分配不同数字区间内的用户至不同的模型；及

- 5 通过所述实时计算后台，调用不同的模型从所述高速缓存区获取不同数字区间内的用户对应的用户特征，根据获取的用户特征进行用户行为预测，并将预测结果实时推荐至用户。

优选地，所述实时推荐系统被所述处理器执行时还用于实现如下步骤：

根据不同模型的预测效果，动态调节不同数字区间的大小；及

- 10 若特定模型的预测效果高于预设阈值，则增大该特定模型所分配的用户数字区间。

优选地，所述实时推荐系统被所述处理器执行时还用于实现如下步骤：

若特定模型的预测效果高于第一预设阈值，则将该特定模型所分配的用户数字区间增大第一比例；及

- 15 若特定模型的预测效果高于第二预设阈值，则将该特定模型所分配的用户数字区间增大第二比例。

优选地，所述模型类继承一个统一的父类，所有模型的实现类通过统一的父类接口进行预测推荐；及

- 20 所述不同数字区间包括第一数字区间和第二数字区间，并将第一数字区间内对应的用户分配至第一模型，第二数字区间内对应的用户分配至第二模型。

优选地，所述实时推荐系统被所述处理器执行时还用于实现如下步骤：

将不同模型的预测效果实时写入所述高速缓存区，并通过特定图表格式将不同模型的预测效果显示于设定的前端界面系统；及

- 25 将每个预测项目的关键数据依据时间顺序统计生成报表，并将生成的报表显示于设定的前端界面系统。

此外，为实现上述目的，本申请还提供一种实时推荐方法，该方法应用于电子设备，所述方法包括：

通过预定编程语言创建模型类，利用该模型类在特定的计算引擎上离线
5 训练模型，并将训练得到的模型托管至实时计算后台；

通过所述实时计算后台，将用户特征存储至高速缓存区；

通过所述实时计算后台，将用户标识映射成指定区间内的数字，并分配
不同数字区间内的用户至不同的模型；及

通过所述实时计算后台，调用不同的模型从所述高速缓存区获取不同数
10 字区间内的用户对应的用户特征，根据获取的用户特征进行用户行为预测，
并将预测结果实时推荐至用户。

优选地，该方法还包括：

根据不同模型的预测效果，动态调节不同数字区间的大小；及

若特定模型的预测效果高于预设阈值，则增大该特定模型所分配的用户
15 数字区间。

优选地，该方法还包括：

若特定模型的预测效果高于第一预设阈值，则将该特定模型所分配的用
户数字区间增大第一比例；及

若特定模型的预测效果高于第二预设阈值，则将该特定模型所分配的用
20 户数字区间增大第二比例。

优选地，所述模型类继承一个统一的父类，所有模型的实现类通过统一
的父类接口进行预测推荐；及

所述不同数字区间包括第一数字区间和第二数字区间，并将第一数字区
间内对应的用户分配至第一模型，第二数字区间内对应的用户分配至第二模
25 型。

进一步地，为实现上述目的，本申请还提供一种计算机可读存储介质，所述计算机可读存储介质存储有实时推荐系统，所述实时推荐系统可被至少一个处理器执行，以使所述至少一个处理器执行如上述的实时推荐方法的步骤。

5

相较于现有技术，本申请所提出的电子设备、实时推荐方法及计算机可读存储介质，通过实时计算后台的运作缩短了预测时间，并利用前端界面系统实现了同时对多个模型进行线上管理。本申请完成了从模型训练到上线到效果评估的闭环，通过创建的模型类可以快速更新迭代模型，同时也可以方便地横向扩展，接入更多的推荐业务。

10

附图说明

图1是本申请电子设备一可选的硬件架构的示意图；
图2是本申请电子设备中实时推荐系统一实施例的程序模块示意图；
图3为本申请实时推荐方法一实施例的实施流程示意图；
图4为本申请输出的模型效果曲线图的示例图。

15

附图标记：

电子设备	2
存储器	21
处理器	22
网络接口	23
实时推荐系统	20
训练模块	201
存储模块	202
分配模块	203

推荐模块	204
流程步骤	S31-S34

本申请目的的实现、功能特点及优点将结合实施例，参照附图做进一步说明。

具体实施方式

5 为了使本申请的目的、技术方案及优点更加清楚明白，以下结合附图及实施例，对本申请进行进一步详细说明。应当理解，此处所描述的具体实施例仅用以解释本申请，并不用于限定本申请。基于本申请中的实施例，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，都属于本申请保护的范围。

10 需要说明的是，在本申请中涉及“第一”、“第二”等的描述仅用于描述目的，而不能理解为指示或暗示其相对重要性或者隐含指明所指示的技术特征的数量。由此，限定有“第一”、“第二”的特征可以明示或者隐含地包括至少一个该特征。另外，各个实施例之间的技术方案可以相互结合，但是必须是以本领域普通技术人员能够实现为基础，当技术方案的结合出现相互矛盾或无法实现时应当认为这种技术方案的结合不存在，也不在本申请要求的

15 保护范围之内。

 进一步需要说明的是，在本文中，术语“包括”、“包含”或者其任何其他变体意在涵盖非排他性的包含，从而使得包括一系列要素的过程、方法、物品或者装置不仅包括那些要素，而且还包括没有明确列出的其他要素，或者

20 是还包括为这种过程、方法、物品或者装置所固有的要素。在没有更多限制的情况下，由语句“包括一个……”限定的要素，并不排除在包括该要素的过程、方法、物品或者装置中还存在另外的相同要素。

首先，本申请提出一种电子设备 2。

参阅图 1 所示，是本申请电子设备 2 一可选的硬件架构的示意图。本实施例中，所述电子设备 2 可包括，但不限于，可通过系统总线相互通信连接存储器 21、处理器 22、网络接口 23。需要指出的是，图 1 仅示出了具有组件 21-23 的电子设备 2，但是应理解的是，并不要求实施所有示出的组件，可以
5 替代的实施更多或者更少的组件。

其中，所述电子设备 2 可以是机架式服务器、刀片式服务器、塔式服务器或机柜式服务器等计算设备，该电子设备 2 可以是独立的服务器，也可以是多个服务器所组成的服务器集群。

所述存储器 21 至少包括一种类型的可读存储介质，所述可读存储介质包
10 括闪存、硬盘、多媒体卡、卡型存储器（例如，SD 或 DX 存储器等）、随机访问存储器（RAM）、静态随机访问存储器（SRAM）、只读存储器（ROM）、电可擦除可编程只读存储器（EEPROM）、可编程只读存储器（PROM）、磁性存储器、磁盘、光盘等。在一些实施例中，所述存储器 21 可以是所述电子设备 2 的内部存储单元，例如该电子设备 2 的硬盘或内存。在另一些实施例中，
15 所述存储器 21 也可以是所述电子设备 2 的外部存储设备，例如该电子设备 2 上配备的插接式硬盘，智能存储卡（Smart Media Card, SMC），安全数字（Secure Digital, SD）卡，闪存卡（Flash Card）等。当然，所述存储器 21 还可以既包括所述电子设备 2 的内部存储单元也包括其外部存储设备。本实施例中，所述存储器 21 通常用于存储安装于所述电子设备 2 的操作系统和各类应用软
20 件，例如所述实时推荐系统 20 的程序代码等。此外，所述存储器 21 还可以用于暂时地存储已经输出或者将要输出的各类数据。

所述处理器 22 在一些实施例中可以是中央处理器（Central Processing Unit, CPU）、控制器、微控制器、微处理器、或其他数据处理芯片。该处理器 22 通常用于控制所述电子设备 2 的总体操作，例如执行与所述电子设备 2
25 进行数据交互或者通信相关的控制和处理等。本实施例中，所述处理器 22 用于运行所述存储器 21 中存储的程序代码或者处理数据，例如运行所述的实时

推荐系统 20 等。

所述网络接口 23 可包括无线网络接口或有线网络接口，该网络接口 23 通常用于在所述电子设备 2 与其他电子设备之间建立通信连接。例如，所述网络接口 23 用于通过网络将所述电子设备 2 与外部数据平台相连，在所述电
5 子设备 2 与外部数据平台之间的建立数据传输通道和通信连接。所述网络可以是企业内部网（Intranet）、互联网（Internet）、全球移动通讯系统（Global System of Mobile communication, GSM）、宽带码分多址（Wideband Code Division Multiple Access, WCDMA）、4G 网络、5G 网络、蓝牙（Bluetooth）、Wi-Fi 等无线或有线网络。

10 至此，已经详细介绍了本申请各个实施例的应用环境和相关设备的硬件结构和功能。下面，将基于上述应用环境和相关设备，提出本申请的各个实施例。

参阅图 2 所示，是本申请电子设备 2 中实时推荐系统 20 一实施例的程序
15 模块图。本实施例中，所述的实时推荐系统 20 可以被分割成一个或多个程序模块，所述一个或者多个程序模块被存储于所述存储器 21 中，并由一个或多个处理器（本实施例中为所述处理器 22）所执行，以完成本申请。例如，在图 2 中，所述的实时推荐系统 20 可以被分割成训练模块 201、存储模块 202、分配模块 203、以及推荐模块 204。本申请所称的程序模块是指能够完成特定
20 功能的一系列计算机程序指令段，比程序更适合于描述所述实时推荐系统 20 在所述电子设备 2 中的执行过程。以下将就各程序模块 201-204 的功能进行详细描述。

所述训练模块 201，用于通过预定编程语言（如 Scala 编程语言）创建模型类，利用该模型类在特定的计算引擎（如 Spark 大数据引擎）上离线训练模
25 型，并将训练得到的模型托管至实时计算后台（如 Java 后台）。在本实施例中，利用 Java 语言与 Scala 语言互通的关系，将 Scala 编程语音创建的模型类训练

得到的模型用 Java 实时计算后台进行托管，从而实现实时预测和推荐，降低预测和推荐的时间。

优选地，在本实施例中，所述模型类继承一个统一的父类，所有模型的实现类通过统一的父类接口进行预测推荐，从而方便横向扩展，接入更多的推荐业务。

举例而言，所述模型类的父类可以定义为如下格式：

```
trait OpmAlgorithmBase extends Serializable{  
  def predict(input:HashMap[String,String]):Any  
}
```

定义好该父类后，所有的模型实现类都继承这个父类，每个模型实现类都需要实现这个预测方法，实时计算后台（如 Java 后台）通过调用这个方法进行实时预测。

所述存储模块 202，用于通过所述实时计算后台，将用户特征（如用户画像特征，可以是大数据量级）存储至高速缓存区（如 Redis 数据库）。在本实施例中，由于用户画像特征缓存到 Redis 数据库，可以大幅缩短用户画像特征查询时间，从而缩短预测时间（单条预测耗时平均 2ms）。

所述分配模块 203，用于通过所述实时计算后台，将用户标识（用户 ID）映射成指定区间内的数字（如哈希成 0-100 之间的数字），并分配不同数字区间内的用户至不同的模型（即训练得到的模型）。例如，将用户标识哈希成 0-100 之间的数字后，分成第一数字区间“0-50”和第二数字区间“51-100”，并将第一数字区间内对应的用户分配至第一模型（如 A 模型），第二数字区间内对应的用户分配至第二模型（如 B 模型）。

所述推荐模块 204，用于通过所述实时计算后台，调用不同的模型从所述

高速缓存区获取不同数字区间内的用户对应的用户特征，根据获取的用户特征进行用户行为预测（如预测用户是否购买产险），并将预测结果实时推荐至用户。

例如，调用第一模型从所述高速缓存区获取第一数字区间内的用户对应的用户特征，根据获取的用户特征针对第一数字区间内的用户进行用户行为预测，并将预测结果实时推荐至第一数字区间内的用户；调用第二模型从所述高速缓存区获取第二数字区间内的用户对应的用户特征，根据获取的用户特征针对第二数字区间内的用户进行用户行为预测，并将预测结果实时推荐至第二数字区间内的用户。

优选地，在其它实施例中，所述实时推荐系统 20 还用于：

根据不同模型的预测效果，动态调节不同数字区间的大小，以便灵活控制每个模型覆盖的预测用户流量，使得预测效果优的模型分配更多的用户。具体而言，在其它实施例中，若特定模型的预测效果高于预设阈值（如 80%），则增大该特定模型所分配的用户数字区间（如增大 20%）。例如，如上举例，若第一模型的预测效果为 85%，则将第一模型所分配的第一数字区间“0-50”增大 20%，得到调节后的第一数字区间“0-60”。相应地，第二模型所分配的第二数字区间“51-100”缩小 20%，得到调节后的第二数字区间“61-100”。

更进一步地，在其它实施例中，若特定模型的预测效果高于第一预设阈值（如 80%），则将该特定模型所分配的用户数字区间增大第一比例（如增大 20%）；若特定模型的预测效果高于第二预设阈值（如 90%），则将该特定模型所分配的用户数字区间增大第二比例（如增大 50%）。

优选地，在其它实施例中，所述实时推荐系统 20 还用于：

将不同模型的预测效果实时写入所述高速缓存区，并通过特定图表格式（参阅图 4 所示的曲线图）将不同模型的预测效果显示于设定的前端界面系

统（如应用程序界面系统），从而可以实时地对比不同模型的优劣，方便同时对多个模型进行线上管理（如 A/B 测试管理等）。

优选地，在其它实施例中，所述实时推荐系统 20 还用于：

- 5 将每个预测项目（或预测业务）的关键数据（如核心用户的信息等）依据时间顺序（如日/周/月/年）统计生成报表，并将生成的报表显示于设定的前端界面系统。

通过上述程序模块 201-204，本申请所提出的实时推荐系统 20，通过实时
10 计算后台的运作缩短了预测时间，并利用前端界面系统实现了同时对多个模型进行线上管理。本申请完成了从模型训练到上线到效果评估的闭环，通过创建的模型类可以快速更新迭代模型，同时也可以方便地横向扩展，接入更多的推荐业务。

- 15 此外，本申请还提出一种实时推荐方法。

参阅图 3 所示，是本申请实时推荐方法一实施例的实施流程示意图。在本实施例中，根据不同的需求，图 3 所示的流程图中的步骤的执行顺序可以改变，某些步骤可以省略。

步骤S31，通过预定编程语言（如Scala编程语言）创建模型类，利用该模
20 型类在特定的计算引擎（如Spark大数据引擎）上离线训练模型，并将训练得到的模型托管至实时计算后台（如Java后台）。在本实施例中，利用Java语言与Scala语言互通的关系，将Scala编程语音创建的模型类训练得到的模型用Java实时计算后台进行托管，从而实现实时预测和推荐，降低预测和推荐的时间。

- 25 优选地，在本实施例中，所述模型类继承一个统一的父类，所有模型的实现类通过统一的父类接口进行预测推荐，从而方便横向扩展，接入更多的

推荐业务。

举例而言，所述模型类的父类可以定义为如下格式：

```
trait OpmAlgorithmBase extends Serializable{  
  def predict(input:HashMap[String,String]):Any  
5    }
```

定义好该父类后，所有的模型实现类都继承这个父类，每个模型实现类都需要实现这个预测方法，实时计算后台（如 Java 后台）通过调用这个方法进行实时预测。

10 步骤S32，通过所述实时计算后台，将用户特征（如用户画像特征，可以是大数据量级）存储至高速缓存区（如Redis数据库）。在本实施例中，由于用户画像特征缓存到Redis数据库，可以大幅缩短用户画像特征查询时间，从而缩短预测时间（单条预测耗时平均2ms）。

15 步骤S33，通过所述实时计算后台，将用户标识（用户ID）映射成指定区间内的数字（如哈希成0-100之间的数字），并分配不同数字区间内的用户至不同的模型（即训练得到的模型）。例如，将用户标识哈希成0-100之间的数字后，分成第一数字区间“0-50”和第二数字区间“51-100”，并将第一数字区间内对应的用户分配至第一模型（如A模型），第二数字区间内对应的用户分配至
20 第二模型（如B模型）。

步骤S34，于所述实时计算后台，调用不同的模型从所述高速缓存区获取不同数字区间内的用户对应的用户特征，根据获取的用户特征进行用户行为预测（如预测用户是否购买产险），并将预测结果实时推荐至用户。

25 例如，调用第一模型从所述高速缓存区获取第一数字区间内的用户对应的用户特征，根据获取的用户特征针对第一数字区间内的用户进行用户行为

预测，并将预测结果实时推荐至第一数字区间内的用户；调用第二模型从所述高速缓存区获取第二数字区间内的用户对应的用户特征，根据获取的用户特征针对第二数字区间内的用户进行用户行为预测，并将预测结果实时推荐至第二数字区间内的用户。

5

优选地，在其它实施例中，所述实时推荐方法还包括如下步骤：

根据不同模型的预测效果，动态调节不同数字区间的大小，以便灵活控制每个模型覆盖的预测用户流量，使得预测效果优的模型分配更多的用户。具体而言，在其它实施例中，若特定模型的预测效果高于预设阈值（如80%），
10 则增大该特定模型所分配的用户数字区间（如增大20%）。例如，如上举例，若第一模型的预测效果为85%，则将第一模型所分配的第一数字区间“0-50”增大20%，得到调节后的第一数字区间“0-60”。相应地，第二模型所分配的第二数字区间“51-100”缩小20%，得到调节后的第二数字区间“61-100”。

更进一步地，在其它实施例中，若特定模型的预测效果高于第一预设阈
15 值（如80%），则将该特定模型所分配的用户数字区间增大第一比例（如增大20%）；若特定模型的预测效果高于第二预设阈值（如90%），则将该特定模型所分配的用户数字区间增大第二比例（如增大50%）。

优选地，在其它实施例中，所述实时推荐方法还包括如下步骤：

20 将不同模型的预测效果实时写入所述高速缓存区，并通过特定图表格式（参阅图4所示的曲线图）将不同模型的预测效果显示于设定的前端界面系统（如应用程序界面系统），从而可以实时地对比不同模型的优劣，方便同时对多个模型进行线上管理（如A/B测试管理等）。

优选地，在其它实施例中，所述实时推荐方法还包括如下步骤：

25 将每个预测项目（或预测业务）的关键数据（如核心用户的信息等）依据时间顺序（如日/周/月/年）统计生成报表，并将生成的报表显示于设定的前

端界面系统。

通过上述步骤 S31-S34 及其它相关步骤，本申请所提出的实时推荐方法，通过实时计算后台的运作缩短了预测时间，并利用前端界面系统实现了同时对多个模型进行线上管理。本申请完成了从模型训练到上线到效果评估的闭环，通过创建的模型类可以快速更新迭代模型，同时也可以方便地横向扩展，接入更多的推荐业务。

进一步地，为实现上述目的，本申请还提供一种计算机可读存储介质（如 ROM/RAM、磁碟、光盘），所述计算机可读存储介质存储有实时推荐系统20，所述实时推荐系统20可被至少一个处理器22执行，以使所述至少一个处理器22执行如上所述的实时推荐方法的步骤。

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上述实施例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来实现，但很多情况下前者是更佳的实施方式。基于这样的理解，本申请的技术方案本质上或者说对现有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，空调器，或者网络设备等）执行本申请各个实施例所述的方法。

以上参照附图说明了本申请的优选实施例，并非因此局限本申请的权利范围。上述本申请实施例序号仅仅为了描述，不代表实施例的优劣。另外，虽然在流程图中示出了逻辑顺序，但是在某些情况下，可以以不同于此处的顺序执行所示出或描述的步骤。

本领域技术人员不脱离本申请的范围和实质，可以有多种变型方案实现本申请，比如作为一个实施例的特征可用于另一实施例而得到又一实施例。

凡是利用本申请说明书及附图内容所作的等效结构或等效流程变换，或直接或间接运用在其他相关的技术领域，均同理包括在本申请的专利保护范围内。

权利要求书

1. 一种电子设备，其特征在于，所述电子设备包括存储器和处理器，所述存储器上存储有可在所述处理器上运行的实时推荐系统，所述实时推荐系统被所述处理器执行时实现如下步骤：

通过预定编程语言创建模型类，利用该模型类在特定的计算引擎上离线训练模型，并将训练得到的模型托管至实时计算后台；

通过所述实时计算后台，将用户特征存储至高速缓存区；

通过所述实时计算后台，将用户标识映射成指定区间内的数字，并分配不同数字区间内的用户至不同的模型；及

通过所述实时计算后台，调用不同的模型从所述高速缓存区获取不同数字区间内的用户对应的用户特征，根据获取的用户特征进行用户行为预测，并将预测结果实时推荐至用户。

2. 如权利要求 1 所述的电子设备，其特征在于，所述实时推荐系统被所述处理器执行时还用于实现如下步骤：

根据不同模型的预测效果，动态调节不同数字区间的大小；及

若特定模型的预测效果高于预设阈值，则增大该特定模型所分配的用户数字区间。

3. 如权利要求 2 所述的电子设备，其特征在于，所述实时推荐系统被所述处理器执行时还用于实现如下步骤：

若特定模型的预测效果高于第一预设阈值，则将该特定模型所分配的用户数字区间增大第一比例；及

若特定模型的预测效果高于第二预设阈值，则将该特定模型所分配的用户数字区间增大第二比例。

4. 如权利要求 2 所述的电子设备，其特征在于，所述模型类继承一个统一的父类，所有模型的实现类通过统一的父类接口进行预测推荐；及

所述不同数字区间包括第一数字区间和第二数字区间，并将第一数字区间内对应的用户分配至第一模型，第二数字区间内对应的用户分配至第二模型。

5. 如权利要求 3 所述的电子设备，其特征在于，所述模型类继承一个统一的父类，所有模型的实现类通过统一的父类接口进行预测推荐；及

所述不同数字区间包括第一数字区间和第二数字区间，并将第一数字区间内对应的用户分配至第一模型，第二数字区间内对应的用户分配至第二模型。

6. 如权利要求 2 所述的电子设备，其特征在于，所述实时推荐系统被所述处理器执行时还用于实现如下步骤：

15 将不同模型的预测效果实时写入所述高速缓存区，并通过特定图表格式将不同模型的预测效果显示于设定的前端界面系统；及

将每个预测项目的关键数据依据时间顺序统计生成报表，并将生成的报表显示于设定的前端界面系统。

20 7. 如权利要求 3 所述的电子设备，其特征在于，所述实时推荐系统被所述处理器执行时还用于实现如下步骤：

将不同模型的预测效果实时写入所述高速缓存区，并通过特定图表格式将不同模型的预测效果显示于设定的前端界面系统；及

25 将每个预测项目的关键数据依据时间顺序统计生成报表，并将生成的报表显示于设定的前端界面系统。

8. 一种实时推荐方法，应用于电子设备，其特征在于，所述方法包括：

通过预定编程语言创建模型类，利用该模型类在特定的计算引擎上离线训练模型，并将训练得到的模型托管至实时计算后台；

通过所述实时计算后台，将用户特征存储至高速缓存区；

5 通过所述实时计算后台，将用户标识映射成指定区间内的数字，并分配不同数字区间内的用户至不同的模型；及

通过所述实时计算后台，调用不同的模型从所述高速缓存区获取不同数字区间内的用户对应的用户特征，根据获取的用户特征进行用户行为预测，并将预测结果实时推荐至用户。

10

9. 如权利要求 8 所述的实时推荐方法，其特征在于，该方法还包括：

根据不同模型的预测效果，动态调节不同数字区间的大小；及

若特定模型的预测效果高于预设阈值，则增大该特定模型所分配的用户数字区间。

15

10. 如权利要求 9 所述的实时推荐方法，其特征在于，该方法还包括：

若特定模型的预测效果高于第一预设阈值，则将该特定模型所分配的用户数字区间增大第一比例；及

20 若特定模型的预测效果高于第二预设阈值，则将该特定模型所分配的用户数字区间增大第二比例。

11. 如权利要求 9 所述的实时推荐方法，其特征在于，所述模型类继承一个统一的父类，所有模型的实现类通过统一的父类接口进行预测推荐；及

25 所述不同数字区间包括第一数字区间和第二数字区间，并将第一数字区间内对应的用户分配至第一模型，第二数字区间内对应的用户分配至第二模型。

12. 如权利要求 10 所述的实时推荐方法，其特征在于，所述模型类继承一个统一的父类，所有模型的实现类通过统一的父类接口进行预测推荐；及

所述不同数字区间包括第一数字区间和第二数字区间，并将第一数字区间内对应的用户分配至第一模型，第二数字区间内对应的用户分配至第二模型。

13. 如权利要求 9 所述的实时推荐方法，其特征在于，该方法还包括：

将不同模型的预测效果实时写入所述高速缓存区，并通过特定图表格式将不同模型的预测效果显示于设定的前端界面系统；及

将每个预测项目的关键数据依据时间顺序统计生成报表，并将生成的报表显示于设定的前端界面系统。

14. 如权利要求 10 所述的实时推荐方法，其特征在于，该方法还包括：

将不同模型的预测效果实时写入所述高速缓存区，并通过特定图表格式将不同模型的预测效果显示于设定的前端界面系统；及

将每个预测项目的关键数据依据时间顺序统计生成报表，并将生成的报表显示于设定的前端界面系统。

15. 一种计算机可读存储介质，所述计算机可读存储介质存储有实时推荐系统，所述实时推荐系统可被至少一个处理器执行，所述实时推荐系统被所述处理器执行时实现如下步骤：

通过预定编程语言创建模型类，利用该模型类在特定的计算引擎上离线训练模型，并将训练得到的模型托管至实时计算后台；

通过所述实时计算后台，将用户特征存储至高速缓存区；

通过所述实时计算后台，将用户标识映射成指定区间内的数字，并分配

不同数字区间内的用户至不同的模型；及

通过所述实时计算后台，调用不同的模型从所述高速缓存区获取不同数字区间内的用户对应的用户特征，根据获取的用户特征进行用户行为预测，并将预测结果实时推荐至用户。

5

16. 如权利要求 15 所述的计算机可读存储介质，其特征在于，所述实时推荐系统被所述处理器执行时还用于实现如下步骤：

根据不同模型的预测效果，动态调节不同数字区间的大小；及

若特定模型的预测效果高于预设阈值，则增大该特定模型所分配的用户数字区间。

10

17. 如权利要求 16 所述的计算机可读存储介质，其特征在于，所述实时推荐系统被所述处理器执行时还用于实现如下步骤：

若特定模型的预测效果高于第一预设阈值，则将该特定模型所分配的用户数字区间增大第一比例；及

15

若特定模型的预测效果高于第二预设阈值，则将该特定模型所分配的用户数字区间增大第二比例。

20

18. 如权利要求 16 所述的计算机可读存储介质，其特征在于，所述模型类继承一个统一的父类，所有模型的实现类通过统一的父类接口进行预测推荐；及

所述不同数字区间包括第一数字区间和第二数字区间，并将第一数字区间内对应的用户分配至第一模型，第二数字区间内对应的用户分配至第二模型。

25

19. 如权利要求 17 所述的计算机可读存储介质，其特征在于，所述模型

类继承一个统一的父类，所有模型的实现类通过统一的父类接口进行预测推荐；及

所述不同数字区间包括第一数字区间和第二数字区间，并将第一数字区间内对应的用户分配至第一模型，第二数字区间内对应的用户分配至第二模型。

20. 如权利要求 16 或 17 所述的计算机可读存储介质，其特征在于，所述实时推荐系统被所述处理器执行时还用于实现如下步骤：

将不同模型的预测效果实时写入所述高速缓存区，并通过特定图表格式将不同模型的预测效果显示于设定的前端界面系统；及

将每个预测项目的关键数据依据时间顺序统计生成报表，并将生成的报表显示于设定的前端界面系统。

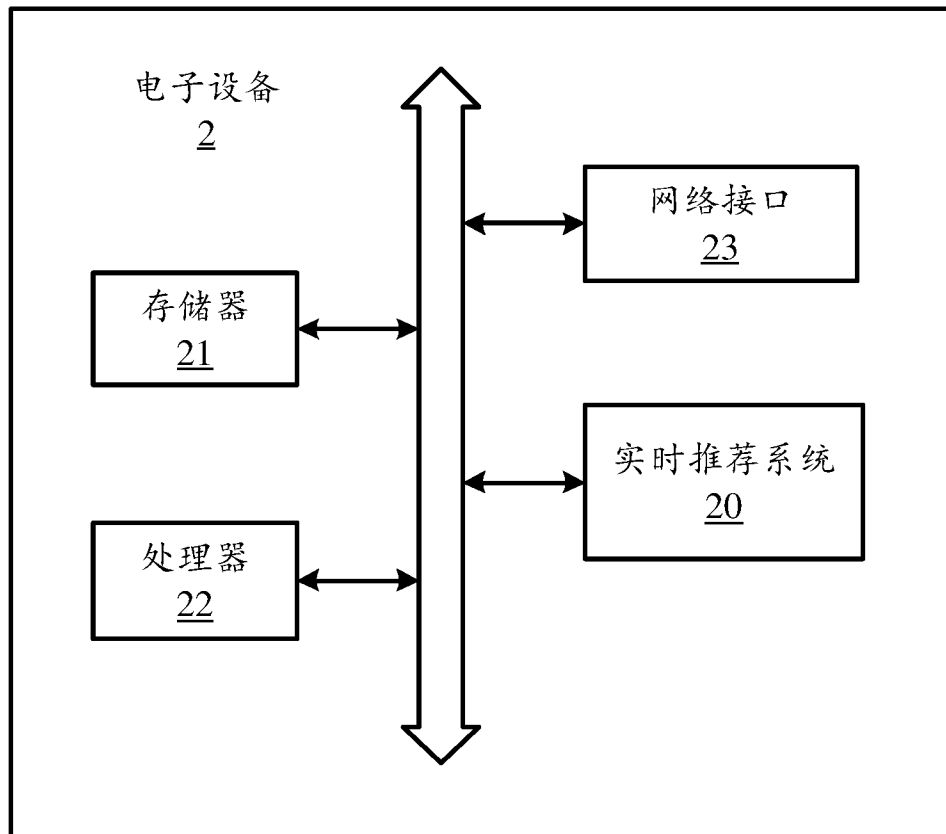


图 1

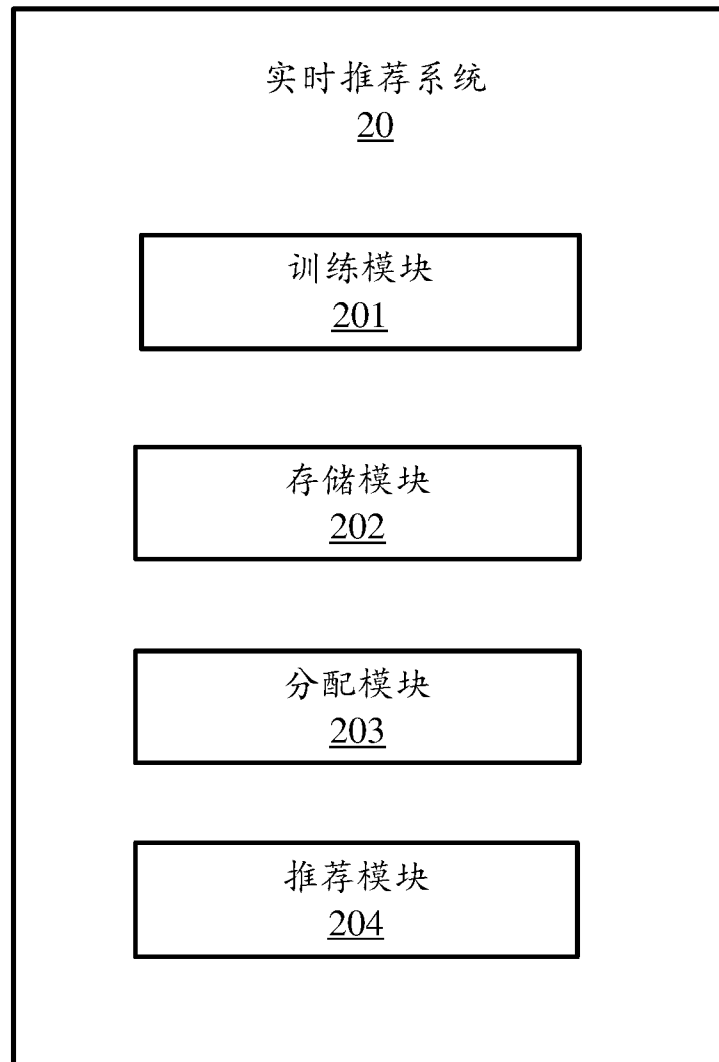


图 2

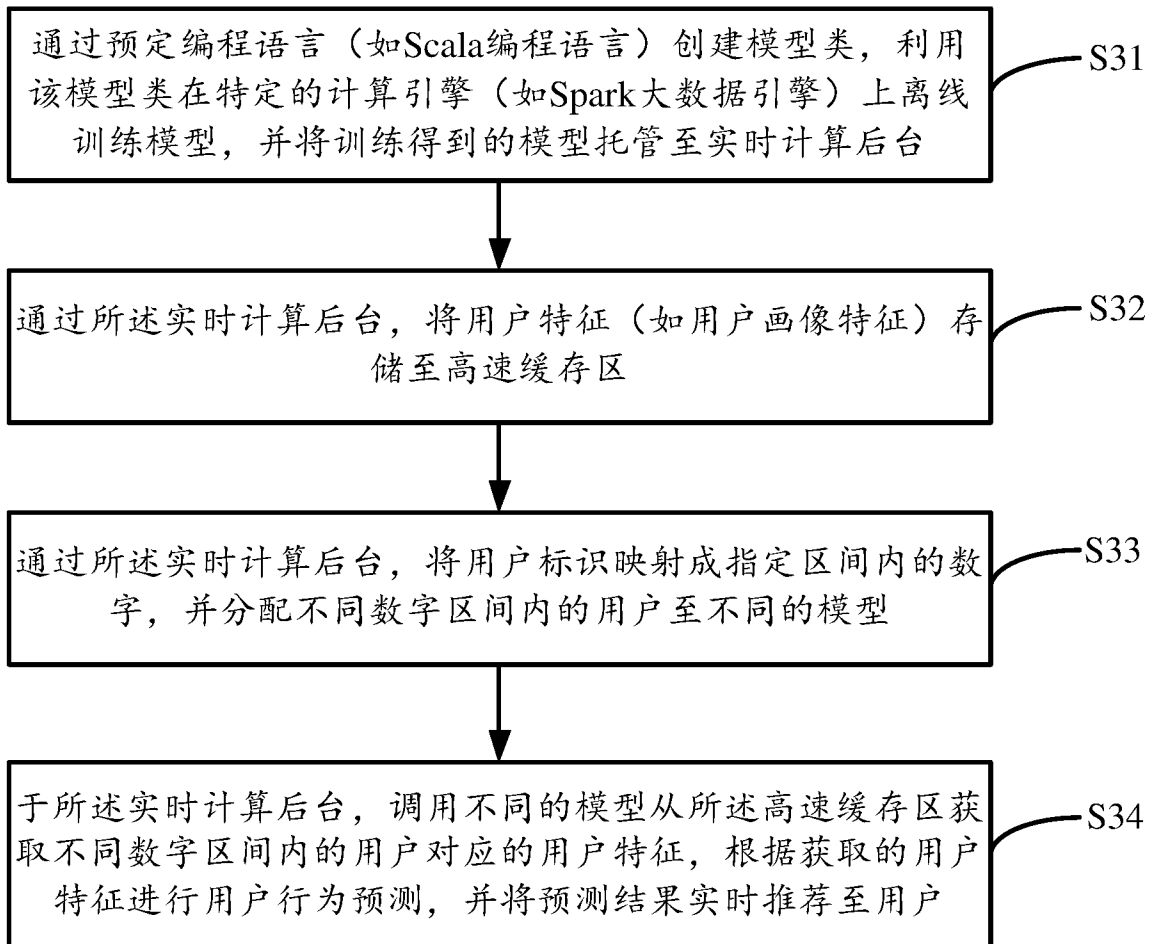


图 3

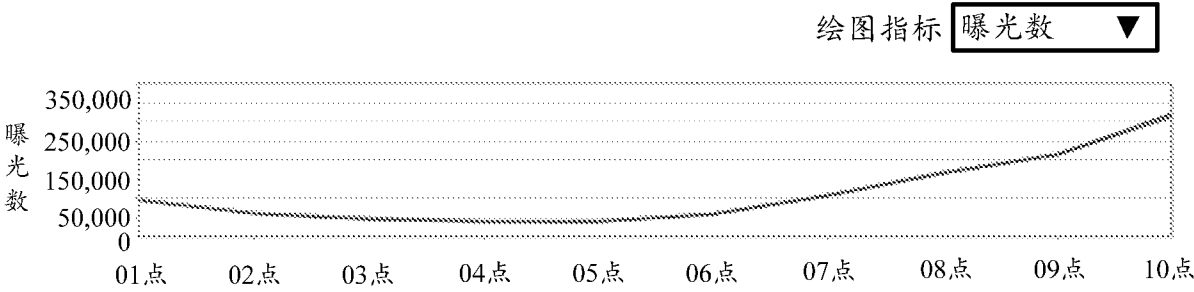


图 4

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2018/076173

A. CLASSIFICATION OF SUBJECT MATTER

G06F 17/30 (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)

CNPAT, CNKI, WPI, EPODOC, IEEE, GOOGLE: 实时, 推荐, 编程语言, 模型, 引擎, 离线, 训练, real w time, commend, scala, model, spark, off w line, training

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	CN 105469263 A (ALIBABA GROUP HOLDING LTD.) 06 April 2016 (06.04.2016), description, paragraphs [0011]-[0019], [0168], [0169] and [0194]-[0196]	1-20
A	CN 107122990 A (GUANGZHOU UC NETWORK TECHNOLOGY CO., LTD.) 01 September 2017 (01.09.2017), entire document	1-20
A	CN 106485562 A (SUNING CORP.) 08 March 2017 (08.03.2017), entire document	1-20
A	CN 106202474 A (TCL CORPORATION) 07 December 2016 (07.12.2016), entire document	1-20
A	US 2015127419 A1 (ORACLE INTERNATIONAL CORPORATION) 07 May 2015 (07.05.2015), entire document	1-20

☐ 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 10 May 2018	Date of mailing of the international search report 30 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 TIAN, Minli Telephone No. (86-10) 53961355

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2018/076173

Patent Documents referred in the Report	Publication Date	Patent Family	Publication Date
CN 105469263 A	06 April 2016	None	
CN 107122990 A	01 September 2017	None	
CN 106485562 A	08 March 2017	None	
CN 106202474 A	07 December 2016	None	
US 2015127419 A1	07 May 2015	JP 2017501477 A	12 January 2017
		WO 2015065590 A1	07 May 2015

国际检索报告

国际申请号

PCT/CN2018/076173

A. 主题的分类

G06F 17/30(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))

CNPAT, CNKI, WPI, EPODOC, IEEE, GOOGLE: 实时, 推荐, 编程语言, 模型, 引擎, 离线, 训练, real w time, commend, scala, model, spark, off w line, training

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 105469263 A (阿里巴巴集团控股有限公司) 2016年 4月 6日 (2016 - 04 - 06) 说明书第[0011]-[0019]、[0168]、[0169]、[0194]-[0196]段	1-20
A	CN 107122990 A (广州优视网络科技有限公司) 2017年 9月 1日 (2017 - 09 - 01) 全文	1-20
A	CN 106485562 A (苏宁云商集团股份有限公司) 2017年 3月 8日 (2017 - 03 - 08) 全文	1-20
A	CN 106202474 A (TCL集团股份有限公司) 2016年 12月 7日 (2016 - 12 - 07) 全文	1-20
A	US 2015127419 A1 (ORACLE INTERNATIONAL CORPORATION) 2015年 5月 7日 (2015 - 05 - 07) 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

2018年 5月 10日

国际检索报告邮寄日期

2018年 5月 30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

田民丽

电话号码 86-(10)-53961355

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2018/076173

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	105469263	A	2016年 4月 6日	无			
CN	107122990	A	2017年 9月 1日	无			
CN	106485562	A	2017年 3月 8日	无			
CN	106202474	A	2016年 12月 7日	无			
US	2015127419	A1	2015年 5月 7日	JP	2017501477	A	2017年 1月 12日
				WO	2015065590	A1	2015年 5月 7日

表 PCT/ISA/210 (同族专利附件) (2015年1月)