一种推荐方法、在线预测模型的发布方法和装置

本申请要求在2020年7月21日提交中国专利局、申请号为202010707254.2的中国专利申请的优先权，该申请的全部内容通过引用结合在本申请中。

技术领域

本申请实施例涉及数据推荐技术，例如涉及一种推荐方法、在线预测模型的发布方法和装置。

背景技术

随着互联网技术的飞速发展，爆炸式增长的信息导致用户想要获取感兴趣的有效内容越来越困难。为了帮助用户快速地获取感兴趣的有效内容，个性化推荐系统应运而生，并在新闻、短视频和音乐等产品中发挥着越来越重要的作用。

在相关技术中，个性化推荐系统一般使用高性能分布式存储系统（如Redis（Remote Dictionary Server，远程字典服务））来存储用户和物料的原始信息，在进行推荐时，个性化推荐系统按需从Redis中获取原始特征，并对原始特征进行预处理后输入给数据模型进行预测打分，然后根据打分结果进行推荐。

然而，这样的推荐系统会存在如下情况：

（1）对于具有海量用户和物料的推荐场景，用户和物料的原始特征消耗大量的高性能分布式存储空间，成本高昂。

（2）随着原始特征越来越多，在高QPS（Queries-per-second，每秒查询率）的场景下，推荐系统访问Redis的耗时增加，且多网络I/O（Input/Output，输入/输出）给带宽带来巨大压力。

（3）在进行预测打分时，数据模型复杂的计算量使得耗时较高。

发明内容

以下是对本文详细描述的主题的概述。本概述并非是为了限制权利要求的保护范围。

本申请提供一种推荐方法及在线预测模型的发布方法和装置，以避免相关

技术中的在线预测模型中原始特征数据量大导致的耗费存储空间、访问耗时以及计算量大等情况。

第一方面，本申请实施例提供了一种推荐方法，所述方法包括：

确定输入至已训练的在线预测模型中的目标用户标识以及多个候选条目标识，其中，所述在线预测模型中存储有多个样本用户标识对应的中间用户特征向量，以及多个样本条目标识对应的中间条目特征向量；

在所述在线预测模型的多个样本用户标识中查找所述目标用户标识，获得所述目标用户标识对应的中间用户特征向量；

在所述在线预测模型的多个样本条目标识中查找每个候选条目标识，获得每个候选条目标识对应的中间条目特征向量；

在所述在线预测模型中，对所述目标用户标识对应的中间用户特征向量以及每个候选条目标识对应的中间条目特征向量进行预测处理，输出每个候选条目标识与所述目标用户标识的匹配度；

根据每个候选条目标识与所述目标用户标识的匹配度，从所述候选条目标识中确定用于推荐至所述目标用户标识的目标条目标识。

第二方面，本申请实施例提供了一种在线预测模型的发布方法，所述方法包括：

获取在线预测模型的模型参数和离线预设模型的模型参数；

获取离线预测模型针对多个样本用户标识生成的中间用户特征向量，以及针对多个样本条目标识生成的中间条目特征向量；

将每个样本用户标识对应的中间用户特征向量、每个样本条目标识对应的中间条目特征向量与所述模型参数存储至预设数据库中；

基于所述预设数据库对所述在线预测模型进行发布。

第三方面，本申请实施例提供了一种推荐装置，所述装置包括：

输入信息确定模块，设置为确定输入至已训练的在线预测模型中的目标用户标识以及多个候选条目标识，其中，所述在线预测模型中存储有多个样本用户标识对应的中间用户特征向量，以及多个样本条目标识对应的中间条目特征向量；

中间用户特征向量查找模块，设置为在所述在线预测模型的多个样本用户标识中查找所述目标用户标识，获得所述目标用户标识对应的中间用户特征向量；

中间条目特征向量查找模块，设置为在所述在线预测模型的多个样本条目标识中查找每个候选条目标识，获得每个候选条目标识对应的中间条目特征向量；

预测处理模块，设置为在所述在线预测模型中，对所述目标用户标识对应的中间用户特征向量以及每个候选条目标识对应的中间条目特征向量进行预测处理，输出每个候选条目标识与所述目标用户标识的匹配度；

目标条目标识确定模块，设置为根据每个候选条目标识与所述目标用户标识的匹配度，从所述候选条目标识中确定用于推荐至所述目标用户标识的目标条目标识。

第四方面，本申请实施例提供了一种在线预测模型的发布装置，所述装置包括：

模型参数获取模块，设置为获取在线预测模型的模型参数和离线预设模型的模型参数；

中间向量获取模块，设置为获取离线预测模型针对多个样本用户标识生成的中间用户特征向量，以及针对多个样本条目标识生成的中间条目特征向量；

存储模块，设置为将每个样本用户标识对应的中间用户特征向量、每个样本条目标识对应的中间条目特征向量与所述模型参数存储至预设数据库中；

模型发布模块，设置为基于所述预设数据库对所述在线预测模型进行发布。

第五方面，本申请实施例提供了一种计算机设备，包括存储器、处理器及存储在存储器上并可在处理器上运行的计算机程序，所述处理器执行所述程序时实现上述的方法。

第六方面，本申请实施例提供了一种计算机可读存储介质，其上存储有计算机程序，该程序被处理器执行时实现上述的方法。

附图说明

图 1 是本申请实施例提供的一种推荐方法实施例的流程图；

图 2 是本申请实施例提供的一种在线预测模型的发布方法实施例的流程图；

图 3 是本申请实施例提供的另一种在线预测模型的发布方法实施例的流程图；

图 4 是本申请实施例提供的预测模型示意图；

图 5 是本申请实施例提供的一种预测模型的训练方法实施例的流程图；

图 6 是本申请实施例提供的标识数据库示意图；

图 7 是本申请实施例提供的用户属性数据库示意图；

图 8 是本申请实施例提供的条目属性数据库示意图；

图 9 是本申请实施例提供的训练数据集示意图；

图 10 是本申请实施例提供的一种在线预测模型的训练和预测场景示例性实施例的流程示意图；

图 11 是本申请实施例提供的一种推荐装置实施例的结构框图；

图 12 是本申请实施例提供的一种在线预测模型的发布装置实施例的结构框图；

图 13 是本申请实施例提供的一种计算机设备的结构示意图。

具体实施方式

下面结合附图和实施例对本申请作详细说明。可以理解的是，此处所描述的示例实施例仅仅用于解释本申请，而非对本申请的限定。另外还需要说明的是，为了便于描述，附图中仅示出了与本申请相关的部分而非全部结构。

图 1 为本申请实施例提供的一种推荐方法实施例的流程图，本实施例可适用于在新闻、短视频和音乐等产品中实现实时推荐的场景，特别适用于用户规模大、下发次数多、时效性强的产品。该方法可以由新闻、短视频和音乐等具有推荐需求的产品中的推荐系统来执行，可以包括如下步骤：

步骤 110、确定输入至已训练的在线预测模型中的目标用户标识以及多个候选条目标识，其中，所述在线预测模型中存储有多个样本用户标识对应的中间用户特征向量，以及多个样本条目标识对应的中间条目特征向量。

在该步骤中，在线预测模型可以为已经发布到线上服务，可供在线服务调用的模型，推荐系统通过调用该在线预测模型来达到既定的推荐目的。

该实施例的在线预测模型中除了包括模型参数以外，还可以缓存一些预先生成的中间张量（tensor）数据。示例性地，上述 tensor 数据至少可以包括：多个样本用户标识（user_id）对应的中间用户特征向量（user_emb），以及多个样本条目标识（item_id）对应的中间条目特征向量（item_emb）。

在一种实施方式中，可以从推荐系统所依附的产品的日志记录中提取历史交互数据，该历史交互数据可以包括发生交互行为的用户标识以及条目标识，

则该用户标识可以作为样本用户标识，该用户标识发生交互行为的条目标识可以作为样本条目标识。

需要说明的是，条目根据产品的不同而不同，例如，在短视频应用程序中，条目可以包括短视频；在音乐应用程序中，条目可以包括音乐、视频等音视频数据；在新闻应用程序中，条目可以包括新闻、图片等；在电商应用程序中，条目可以包括商品、广告等。

在该实施例中，在线预测模型的输入可以包括目标用户标识以及多个候选条目标识。

步骤120、在所述在线预测模型的多个样本用户标识中查找所述目标用户标识，获得所述目标用户标识对应的中间用户特征向量。

在一种实施例中，在在线预测模型中可以将多个样本用户标识与对应的中间用户特征向量关联存储，例如，可以以键值对（Key-Value）的方式存储两者，以样本用户标识作为键Key，以中间用户特征向量作为对应的值Value，若中间用户特征向量较多，则以数组的形式来存储，得到的存储数据可以为“user_id:{user_emb 1，user_emb 2，user_emb 3，…}”。

则在步骤120中，当在线预测模型接收到输入的目标用户标识以后，可以在缓存的多个样本用户标识中查找该目标用户标识，例如，在缓存的Key列表中查找该目标用户标识。如果查找成功，则可以获得该查找成功的Key对应的Value值作为该目标用户标识对应的中间用户特征向量。

在其他实施例中，若在在线预测模型缓存的多个样本用户标识中查找不到该目标用户标识，例如，针对目标用户标识为新注册用户，在当前产品中尚未发生交互（如视频交互等）的情况，则可以将该目标用户标识对应的中间用户特征向量设置为第一预设向量。其中，第一预设向量可以按需设置，本实施例对此不作限制，例如可以将对应的中间用户特征向量设置为0向量。

步骤130、在所述在线预测模型的多个样本条目标识中查找每个候选条目标识，获得每个候选条目标识对应的中间条目特征向量。

在一种实施例中，在在线预测模型中可以将每个样本条目标识与对应的中间条目特征向量关联存储，例如，可以以键值对（Key-Value）的方式存储两者，以样本条目标识作为键Key，以中间条目特征向量作为对应的值Value，若中间条目特征向量较多，则以数组的形式来存储，得到的存储数据可以为“item_id:{item _emb 1，item _emb 2，item _emb 3，…}”。

则在步骤 130 中，当在线预测模型接收到多个候选条目标识以后，可以在缓存的多个样本条目标识中查找每个候选条目标识，例如，在缓存的 Key 列表中查找每个候选条目标识。如果查找成功，则可以获得该查找成功的 Key 对应的 Value 值作为该候选条目标识对应的中间条目特征向量。

在其他实施例中，若在在线预测模型缓存的多个样本条目标识中查找不到某个候选条目标识，例如，针对某个短视频为新上线的短视频，在当前产品中尚未发生交互（如视频交互等）的情况，则可以将该候选条目标识对应的中间条目特征向量设置为第二预设向量。其中，第二预设向量可以按需设置，本实施例对此不作限制，例如可以将对应的中间条目特征向量设置为 0 向量。

步骤 140、在所述在线预测模型中，对所述目标用户标识对应的中间用户特征向量以及每个候选条目标识对应的中间条目特征向量进行预测处理，输出每个候选条目标识与所述目标用户标识的匹配度。

在该步骤中，当获得目标用户标识对应的中间用户特征向量以及每个候选条目标识对应的中间条目特征向量以后，则在线预测模型可以按照预设的预测处理逻辑计算该中间用户特征向量与每个中间条目特征向量的匹配度。示例性地，该匹配度可以为匹配分数，在线预测模型可以针对目标用户标识对应的中间用户特征向量，根据每个候选条目标识对应的中间条目特征向量，对每个候选条目的预估 CTR（Click-Through-Rate，点击通过率，指下发条目的点击到达率，即该条目的实际点击次数除以条目展现量）进行打分，其中分数越高，说明该候选条目标识与该目标用户标识越匹配，反之，分数越低，说明该候选条目标识与该目标用户标识越不匹配。

需要说明的是，上述的预测处理逻辑可以采用基于在线预测模型的常用技术手段实现，本实施例对此不作限定。

步骤 150、根据每个候选条目标识与所述目标用户标识的匹配度，从所述候选条目标识中确定用于推荐至所述目标用户标识的目标条目标识。

在一种实施方式中，可以按照匹配度对每个候选条目标识进行排序，并将排序在前的、分数较高的若干个候选条目标识作为目标条目标识。其中，本实施例对目标条目标识的数量不作限制，可以根据具体的需求由用户或开发人员设定。

当确定目标条目标识以后，可以将该目标条目标识对应的目标条目推荐至目标用户标识中。

在本实施例中，在进行条目推荐时，在线预测模型根据输入的目标用户标识以及多个候选条目标识，可以直接从模型的缓存中查找对应的中间用户特征向量以及中间条目特征向量，省略了获取目标用户标识以及多个候选条目标识对应的原始特征并对原始特征进行预处理的过程，从而有效地降低了在线预测模型的计算量，减少机器成本的同时，降低了模型的预估耗时和响应延时，提升了推荐系统中模型预测的性能。

另外，由于在在线预测模型中预先存储有多个样本用户标识对应的中间用户特征向量，以及多个样本条目标识对应的中间条目特征向量，这样在线预测模型在进行推荐时，针对同一用户的不同推荐场景，或者同一条目的不同推荐场景，可以直接查找对应的中间用户特征向量或者中间条目特征向量，无需重复计算中间向量，进一步降低了在线预测模型的计算量。

图 2 为本申请实施例提供的一种在线预测模型的发布方法实施例的流程图，本实施例可适用于线上发布在线预测模型的场景，该方法可以由发布装置来执行，可以包括如下步骤：

步骤 210、获取在线预测模型的模型参数和离线预测模型的模型参数。

模型参数是机器学习算法的关键，是模型内部的变量，在实现中，模型参数可以通过模型代码使用训练数据训练后得到。其值可以根据数据进行估计，模型参数的值定义了可使用的模型。

示例性地，在线预测模型的模型参数可以包括但不限于：全连接层的权重、BN（Batch Normalization）层的均值和方差等网络参数和模型向量（embedding）参数。

步骤 220、获取离线预测模型针对多个样本用户标识生成的中间用户特征向量，以及针对多个样本条目标识生成的中间条目特征向量。

在该实施例中，可以预训练离线预测模型，该离线预测模型设置为针对多个样本用户标识生成对应的中间用户特征向量，以及针对多个样本条目标识生成对应中间条目特征向量。

在一种实施方式中，离线预测模型可以包括用户离线预测模型以及条目离线预测模型，则步骤 220 可以包括如下步骤：

步骤 220-1，分别将每个样本用户标识及对应的用户属性信息输入至用户离线预测模型，并获取所述用户离线预测模型输出的该样本用户标识对应的中间

用户特征向量。

在该步骤中，用户离线预测模型可以是一种深度神经网络模型，用户离线预测模型的输入包括样本用户标识及对应的用户属性信息，如年龄、性别、用户统计特征、用户历史行为等，该模型的输出则为该样本用户标识对应的中间用户特征向量。

在一种实施方式中，样本用户标识对应的用户属性信息可以来源于当前推荐系统所依附的产品对应的用户数据库，该用户数据库示例性地可以用于存储所有已注册的用户标识以及该用户标识对应的其他属性信息。

步骤 220-2，分别将每个样本条目标识及对应的条目属性信息输入至条目离线预测模型，并获取所述条目离线预测模型输出的该样本条目标识对应的中间条目特征向量。

在该步骤中，条目离线预测模型也可以是一种深度神经网络模型，条目离线预测模型的输入包括样本条目标识及对应的条目属性信息，例如，在短视频推荐场景中，样本条目标识可以为短视频 ID，对应的条目属性信息可以包括如视频标签、拍客 id、视频统计信息、国家信息等，该模型的输出则为该样本条目标识对应的中间条目特征向量。

在一种实施方式中，样本条目标识对应的条目属性信息可以来源于当前推荐系统所依附的产品对应的条目数据库，该条目数据库可以用于存储所有已上线的条目的条目标识以及该条目标识对应的条目属性信息。

步骤 230、将每个样本用户标识对应的中间用户特征向量、每个样本条目标识对应的中间条目特征向量与所述模型参数存储至预设数据库中。

在该步骤中，当获得每个样本用户标识对应的中间用户特征向量以及各样本条目标识对应的中间条目特征向量以后，可以将每个样本用户标识对应的中间用户特征向量、每个样本条目标识对应的中间条目特征向量与在线预测模型的模型参数一起存储在模型的预设数据库中。这样，多个样本用户标识对应的中间用户特征向量以及多个样本条目标识对应的中间条目特征向量可以作为模型参数的一部分存储在模型中。

步骤 240、基于所述预设数据库对所述在线预测模型进行发布。

在该步骤中，可以将预设数据库中缓存的中间 tensor 数据（包括上述的中间用户特征向量和中间条目特征向量）随着模型参数一起发布到线上服务器，供在线服务调用。

在本实施例中，在发布在线预测模型时，可以将中间用户特征向量和中间条目特征向量随着模型参数一起发布，使得在线预测模型在进行预测处理时无需重复计算中间用户特征向量和中间条目特征向量，从而降低了在线预测模型的计算量，提高了在线预测模型的预估效率。

图 3 为本申请实施例提供的另一种在线预测模型的发布方法实施例的流程图，本实施例在上述实施例的基础上，对获取在线预测模型的模型参数和离线预设模型的模型参数部分进行了示例性说明，可以包括如下步骤：

步骤 310、训练预测模型。

示例性地，预测模型的训练数据可以包括多个样本用户标识及其对应的用户属性信息，以及，多个样本条目标识及其对应的条目属性信息。例如，预测模型的每条训练数据记录至少可以包括：样本用户标识、样本条目标识、用户属性信息、条目属性信息等。

在一种实施方式中，可以采用 Batch（批）训练的方式对训练数据进行训练，生成预测模型。另外，在模型训练的同时，可以记录样本里出现过的样本用户标识和样本条目标识，以供后续流程使用。

步骤 320、对所述预测模型按照预设裁剪规则进行裁剪，获得在线预测模型的模型参数以及离线预测模型的模型参数。

在一种实施例中，如图 4 所示，训练完成的预测模型可以包括在线预测模型（online_predict_network）和离线预测模型（offline_predict_network）两部分，离线预测模型设置为离线生成多个样本的中间 tensor 数据；在线预测模型设置为实时线上预测，输出匹配分数（rank_scores）。

在生成预测模型以后，可以按照预设裁剪规则对预测模型进行裁剪，得到在线预测模型和离线预测模型。其中，裁剪规则可以由开发人员按需设定，例如在本实施例中，可以按照功能将线上 inference（预测）的必要部分归为线上网络的部分，将剩下的部分归为离线预测模型。

步骤 330、获取所述离线预测模型的模型参数对应的离线预测模型针对多个样本用户标识生成的中间用户特征向量，以及针对多个样本条目标识生成的中间条目特征向量。

在一种实施方式中，如图 4 所示，离线预测模型可以包括用户离线预测模型（user_side_network）以及条目离线预测模型（item_side_network），则步骤

330 可以包括如下步骤：

步骤 330-1，分别将每个样本用户标识及对应的用户属性信息输入至用户离线预测模型，并获取所述用户离线预测模型输出的该样本用户标识对应的中间用户特征向量。

在该步骤中，用户离线预测模型可以是一种深度神经网络模型，如图 4 所示，用户离线预测模型的输入包括样本用户标识（user_id）及对应的用户属性信息，如年龄（age）、性别（gender）、用户统计特征（user_dense）、用户历史行为（user_history）等，该模型的输出则为该样本用户标识对应的中间用户特征向量，如 user_emb1、user_emb2、user_emb3 等 tensor 数据。

在一种实施方式中，样本用户标识对应的用户属性信息可以来源于当前推荐系统所依附的产品对应的用户数据库。

步骤 330-2，分别将每个样本条目标识及对应的条目属性信息输入至条目离线预测模型，并获取所述条目离线预测模型输出的该样本条目标识对应的中间条目特征向量。

条目离线预测模型也可以是一种深度神经网络模型，条目离线预测模型的输入包括样本条目标识及对应的条目属性信息，例如，在短视频推荐场景中，如图 4 所示，样本条目标识可以为短视频 ID（即图 4 的 vid），对应的条目属性信息可以包括如视频标签（tag）、拍客 id（即图 4 的 pid）、视频统计信息（item_dense）、国家信息（country）等，该模型的输出则为该样本条目标识对应的中间条目特征向量，如 item_emb1、item_emb2、item_emb3 等 tensor 数据。

步骤 340、将每个样本用户标识对应的中间用户特征向量、每个样本条目标识对应的中间条目特征向量与所述在线预测模型的模型参数存储至预设数据库中。

在该步骤中，需要建库缓存从离线预测模型中获取的 tensor 数据，如图 4 所示，将离线预测模型输出的 user_emb1、user_emb2、user_emb3、item_emb1、item_emb2、item_emb3 等 tensor 数据与模型参数一起存储在在线预测模型的预设数据库中。

步骤 350、基于所述预设数据库对所述在线预测模型进行发布。

在本实施例中，将预测模型分成在线预测模型和离线预测模型两部分，并将在线预测模型从预测模型中裁剪出来发布到线上服务器，相比于发布整个预测模型，可以降低在线预测模型的网络结构的复杂度，减少机器成本，且为更

复杂庞大的模型上线提供可能。

图 5 为本申请实施例提供的一种预测模型的训练方法实施例的流程图，本实施例在上述实施例的基础上，预测模型的训练过程进行了示例性说明，可以包括如下步骤：

步骤 510、将预先建立的标识数据库、用户属性数据库以及条目属性数据库加载进内存，其中，所述标识数据库用于存储多个样本用户标识及多个样本用户标识发生交互的样本条目标识；所述用户属性数据库用于存储多个样本用户标识对应的用户属性信息；所述条目属性数据库用于存储多个样本条目标识对应的条目属性信息。

在该实施例中，可以将样本数据分成三部分进行存储，用于存储的数据库包括标识数据库、用户属性数据库以及条目属性数据库。

其中，标识数据库用于存储多个样本用户标识及多个样本用户标识发生交互的样本条目标识，例如，如图 6 所示，标识数据库中存储的是多个样本用户标识 user_id 以及该 user_id 产生交互行为的样本条目标识 item_id。在一种实现中，可以从推荐系统所依附的产品的日志记录中提取历史交互数据，该历史交互数据可以包括发生交互行为的用户标识以及条目标识，则该用户标识可以作为样本用户标识，该用户标识发生交互行为的条目标识可以作为样本条目标识。然后，将每个样本用户标识以及该样本用户标识产生交互行为的样本条目标识关联存储到标识数据库中。

用户属性数据库用于存储多个样本用户标识对应的用户属性信息，例如，如图 7 所示，用户属性数据库中存储的是 user_id 对应的用户属性信息，如年龄 age、性别 gender 等。在一种实现中，可以整合标识数据库中的样本用户标识，例如在图 6 中出现两个 user_id 为 110 的标识，则整合时在用户属性数据库中记录一个 user_id 为 110 的标识的数据记录则可，如图 7 第一条记录所示。然后从当前产品的用户数据库中查找每个样本用户标识对应的用户属性信息存入用户属性数据库中。

条目属性数据库用于存储多个样本条目标识对应的条目属性信息，例如，如图 8 所示，条目属性数据库中存储的是 item_id 对应的条目属性信息，如标签 tag、国家 country 等。在一种实现中，可以整合标识数据库中相同的样本条目标识，然后从当前产品的条目数据库中查找每个样本条目标识对应的条目属性信

息存入条目属性数据库中。

在进行模型训练时，可以先将标识数据库、用户属性数据库以及条目属性数据库加载进内存，以提高数据读取的效率。

步骤 520、根据所述标识数据库中记录的样本用户标识及样本条目标识，分别从所述用户属性数据库与所述条目属性数据库中查找出对应的用户属性信息以及条目属性信息，与所述样本用户标识及样本条目标识组织成训练数据。

在该步骤中，针对标识数据库中的每条样本记录，可以从用户属性数据库与条目属性数据库中查找出对应的用户属性信息以及条目属性信息进行特征补全，得到训练数据。例如，针对图 6 的标识数据库中的第一条样本记录“user_id=110，item_id=6010”，从图 7 的用户属性数据库中查找出 user_id=110 对应的数据记录，获得该数据记录中的用户属性信息“age=16, gender=男”，并从图 8 的样本属性数据库中查找出 item_id=6010 对应的数据记录，获得该数据记录中的样本属性信息“tag=运动，室外，country=IN”，将以上获得的数据作为“user_id=110，item_id=6010”的特征进行补全，得到训练数据，如此类推，遍历图 6 中的所有数据记录，最终得到的训练数据集如图 9 所示。

步骤 530、按照预设的模型训练规则对所述训练数据进行模型训练，生成预测模型。

在一种实现方式中，可以采用 Batch 训练的方式对训练数据进行训练，得到预测模型。

在本实施例中，将样本数据分成三个部分，采用三个数据库进行存储，分别为用于存储多个样本用户标识及多个样本用户标识发生交互的样本条目标识的标识数据库，用于存储多个样本用户标识对应的用户属性信息的用户属性数据库，以及用于存储多个样本条目标识对应的条目属性信息的条目属性数据库。当需要训练模型时，根据标识数据库中记录的样本用户标识及样本条目标识，分别从用户属性数据库与条目属性数据库中查找出对应的用户属性信息以及条目属性信息，与样本用户标识及样本条目标识组织成训练数据，通过这样的方式可以复用用户属性数据库与条目属性数据库中的数据，达到压缩样本数据的目的。

图 10 为本申请实施例提供的一种在线预测模型的训练和预测场景示例性实施例的流程示意图，在该流程示意图中，包括如下几个流程：样本数据压缩流

程、模型训练流程、模型建库流程以及线上排序流程。

在样本数据压缩流程中，修改了常规的样本数据的存储模式，将样本数据分成三部分：标识数据库、用户属性数据库以及条目属性数据库。标识数据库用于存储 user_id 和 item_id，用户属性信息和条目属性信息分别存储于用户属性数据库和条目属性数据库，将多个样本共用的特征部分抽离成单独的特征文件（即用户属性数据库以及条目属性数据库），通过这样的方式复用特征数据达到训练数据压缩的目的。

在模型训练流程中，在训练时将样本共用的特征数据加载在内存，训练时首先根据标识数据库中存储的 user_id 和 item_id，从内存中读取相应的属性特征对样本进行特征补全，然后使用 Batch 训练的方式对模型进行训练和更新。在模型训练的同时，还可以记录样本里出现过的 user_id 和 item_id，以供后续流程使用。

在模型建库流程中，加载上一阶段训练得到的预测模型，同时输入上一阶段记录的样本里出现过的 user_id 和 item_id，从内存里读取相应的属性特征对样本进行特征补全，然后使用预测模型对样本进行预测。在预估过程中，将离线预测模型中的部分中间 tensor 缓存到在线预测模型，和在线预测模型本来的模型参数一同存储。

在线上排序流程中，将已经缓存中间 tensor 数据的在线预测模型发布到线上服务，供在线服务调用。在线预测模型被调用后对候选条目进行打分时，无需再读取和预处理用户和候选条目的原始特征，输入目标用户标识和候选条目标识给在线预测模型即可，在线预测模型根据目标用户标识和候选条目标识在缓存的 tensor 数据里查找出对应的 tensor 数据，根据查找出的 tensor 数据进行预测。

通过本实施例，可以改善实时线上推荐系统中模型预测的性能，能够降低线上模型预估耗时的同时，显著减少机器成本。

图 11 为本申请实施例提供的一种推荐装置实施例的结构框图，该推荐装置可以包括：

输入信息确定模块 1110，设置为确定输入至已训练的在线预测模型中的目标用户标识以及多个候选条目标识，其中，所述在线预测模型中存储有多个样本用户标识对应的中间用户特征向量，以及多个样本条目标识对应的中间条目

特征向量；

中间用户特征向量查找模块 1120，设置为在所述在线预测模型的多个样本用户标识中查找所述目标用户标识，获得所述目标用户标识对应的中间用户特征向量；

中间条目特征向量查找模块 1130，设置为在所述在线预测模型的多个样本条目标识中查找每个候选条目标识，获得每个候选条目标识对应的中间条目特征向量；

预测处理模块 1140，设置为在所述在线预测模型中，对所述目标用户标识对应的中间用户特征向量以及每个候选条目标识对应的中间条目特征向量进行预测处理，输出每个候选条目标识与所述目标用户标识的匹配度；

目标条目标识确定模块 1150，设置为根据每个候选条目标识与所述目标用户标识的匹配度，从所述候选条目标识中确定用于推荐至所述目标用户标识的目标条目标识。

在本申请实施例的一种可能的实施例中，所述推荐装置还可以包括：

第一预设向量设置模块，设置为若在所述在线预测模型的多个样本用户标识中查找不到所述目标用户标识，则将所述目标用户标识对应的中间用户特征向量设置为第一预设向量；

第二预设向量设置模块，设置为若在所述在线预测模型的多个样本条目标识中查找不到所述候选条目标识，则将该候选条目标识对应的中间条目特征向量设置为第二预设向量。

图 12 为本申请实施例提供的一种在线预测模型的发布装置实施例的结构框图，该装置可以包括：

模型参数获取模块 1210，设置为获取在线预测模型的模型参数和离线预设模型的模型参数；

中间向量获取模块 1220，设置为获取离线预测模型针对多个样本用户标识生成的中间用户特征向量，以及针对多个样本条目标识生成的中间条目特征向量；

存储模块 1230，设置为将每个样本用户标识对应的中间用户特征向量、每个样本条目标识对应的中间条目特征向量与所述模型参数存储至预设数据库中；

模型发布模块 1240，设置为基于所述预设数据库对所述在线预测模型进行

发布。

在本申请实施例的一种可能的实施例中，所述模型参数获取模块 1210 可以包括：

预测模型训练子模块，设置为训练预测模型；

裁剪子模块，设置为对所述预测模型按照预设裁剪规则进行裁剪，获得在线预测模型的模型参数以及离线预测模型的模型参数。

在本申请实施例的一种可能的实施例中，所述预测模型训练子模块包括：

内存加载单元，设置为将预先建立的标识数据库、用户属性数据库以及条目属性数据库加载进内存，其中，所述标识数据库用于存储多个样本用户标识及多个样本用户标识发生交互的样本条目标识；所述用户属性数据库用于存储多个样本用户标识对应的用户属性信息；所述条目属性数据库用于存储多个样本条目标识对应的条目属性信息；

训练数据组织单元，设置为根据所述标识数据库中记录的样本用户标识及样本条目标识，分别从所述用户属性数据库与所述条目属性数据库中查找出对应的用户属性信息以及条目属性信息，与所述样本用户标识及样本条目标识组织成训练数据；

模型训练单元，设置为按照预设的模型训练规则对所述训练数据进行模型训练，生成预测模型。

在本申请实施例的一种可能的实施例中，所述离线预测模型包括用户离线预测模型以及条目离线预测模型，所述中间向量获取模块 1220 包括：

中间用户特征向量获取子模块，设置为分别将每个样本用户标识及对应的用户属性信息输入至用户离线预测模型，并获取所述用户离线预测模型输出的该样本用户标识对应的中间用户特征向量；

中间条目特征向量获取子模块，设置为分别将每个样本条目标识及对应的条目属性信息输入至条目离线预测模型，并获取所述条目离线预测模型输出的该样本条目标识对应的中间条目特征向量。

需要说明的是，本申请实施例所提供的上述装置可执行本申请任意实施例所提供的方法，具备执行方法相应的功能模块和有益效果。

图 13 为本申请一实施例提供的一种计算机设备的结构示意图，如图 13 所示，该设备/终端/服务器包括处理器 130、存储器 131、输入装置 132 和输出装

置 133；计算机设备中处理器 130 的数量可以是一个或多个，图 13 中以一个处理器 130 为例；计算机设备中的处理器 130、存储器 131、输入装置 132 和输出装置 133 可以通过总线或其他方式连接，图 13 中以通过总线连接为例。

存储器 131 作为一种计算机可读存储介质，可设置为存储软件程序、计算机可执行程序以及模块，如本申请实施例中的推荐方法或在线预测模型的发布方法对应的程序指令/模块（例如，推荐装置中的输入信息确定模块 1110、中间用户特征向量查找模块 1120、中间条目特征向量查找模块 1130、预测处理模块 1140 和目标条目标识确定模块 1150；或者，在线预测模型的发布装置中的模型参数获取模块 1210、中间向量获取模块 1220、存储模块 1230 和模型发布模块 1240）。处理器 130 通过运行存储在存储器 131 中的软件程序、指令以及模块，从而执行计算机设备的多种功能应用以及数据处理，即实现上述的推荐方法或在线预测模型的发布方法。

存储器 131 可主要包括存储程序区和存储数据区，其中，存储程序区可存储操作系统、至少一个功能所需的应用程序；存储数据区可存储根据终端的使用所创建的数据等。此外，存储器 131 可以包括高速随机存取存储器，还可以包括非易失性存储器，例如至少一个磁盘存储器件、闪存器件、或其他非易失性固态存储器件。在一些实例中，存储器 131 可包括相对于处理器 130 远程设置的存储器，这些远程存储器可以通过网络连接至计算机设备。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。

输入装置 132 可设置为接收输入的数字或字符信息，以及产生与计算机设备的用户设置以及功能控制有关的键信号输入。输出装置 133 可包括显示屏等显示设备。

实施例九

本申请实施例九还提供一种包含计算机可执行指令的存储介质，所述计算机可执行指令在由计算机处理器执行时用于执行前述实施例中任一实施例中的方法。

通过以上关于实施方式的描述，所属领域的技术人员可以清楚地了解到，本申请可借助软件及必需的通用硬件来实现，当然也可以通过硬件实现。基于这样的理解，本申请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，该计算机软件产品可以存储在计算机可读存储介

质中，如计算机的软盘、只读存储器（Read-Only Memory， ROM）、随机存取存储器（Random Access Memory， RAM）、闪存（FLASH）、硬盘或光盘等，包括若干指令用以使得一台计算机设备（可以是个人计算机，服务器，或者网络设备等）执行本申请多个实施例所述的方法。

值得注意的是，上述搜索装置的实施例中，所包括的多个单元和模块只是按照功能逻辑进行划分的，但并不局限于上述的划分，只要能够实现相应的功能即可；另外，多功能单元的具体名称也只是为了便于相互区分，并不用于限制本申请的保护范围。

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

权 利 要 求 书

1、一种推荐方法，包括：

确定输入至已训练的在线预测模型中的目标用户标识以及多个候选条目标识，其中，所述在线预测模型中存储有多个样本用户标识对应的中间用户特征向量，以及多个样本条目标识对应的中间条目特征向量；

在所述在线预测模型的多个样本用户标识中查找所述目标用户标识，获得所述目标用户标识对应的中间用户特征向量；

在所述在线预测模型的多个样本条目标识中查找每个候选条目标识，获得每个候选条目标识对应的中间条目特征向量；

在所述在线预测模型中，对所述目标用户标识对应的中间用户特征向量以及每个候选条目标识对应的中间条目特征向量进行预测处理，输出每个候选条目标识与所述目标用户标识的匹配度；

根据每个候选条目标识与所述目标用户标识的匹配度，从所述候选条目标识中确定用于推荐至所述目标用户标识的目标条目标识。

2、根据权利要求 1 所述的推荐方法，所述在所述在线预测模型中，对所述目标用户标识对应的中间用户特征向量以及每个候选条目标识对应的中间条目特征向量进行预测处理，输出每个候选条目标识与所述目标用户标识的匹配度之前，所述方法还包括：

响应于确定在所述在线预测模型的多个样本用户标识中查找不到所述目标用户标识，将所述目标用户标识对应的中间用户特征向量设置为第一预设向量；

或者，

响应于确定在所述在线预测模型的多个样本条目标识中查找不到所述候选条目标识，将所述候选条目标识对应的中间条目特征向量设置为第二预设向量。

3、一种在线预测模型的发布方法，包括：

获取在线预测模型的模型参数和离线预设模型的模型参数；

获取离线预测模型针对多个样本用户标识生成的中间用户特征向量，以及针对多个样本条目标识生成的中间条目特征向量；

将每个样本用户标识对应的中间用户特征向量、每个样本条目标识对应的中间条目特征向量与所述模型参数存储至预设数据库中；

基于所述预设数据库对所述在线预测模型进行发布。

4、根据权利要求 3 所述的在线预测模型的发布方法，其中，所述获取在线

预测模型的模型参数和离线预设模型的模型参数，包括：

训练预测模型；

对所述预测模型按照预设裁剪规则进行裁剪，获得在线预测模型的模型参数以及离线预测模型的模型参数。

5、根据权利要求 4 所述的在线预测模型的发布方法，其中，所述训练预测模型包括：

将预先建立的标识数据库、用户属性数据库以及条目属性数据库加载进内存，其中，所述标识数据库用于存储多个样本用户标识及所述多个样本用户标识发生交互的样本条目标识；所述用户属性数据库用于存储所述多个样本用户标识对应的用户属性信息；所述条目属性数据库用于存储所述多个样本条目标识对应的条目属性信息；

根据所述标识数据库中记录的样本用户标识及样本条目标识，分别从所述用户属性数据库与所述条目属性数据库中查找出对应的用户属性信息以及条目属性信息，与所述样本用户标识及样本条目标识组织成训练数据；

按照预设的模型训练规则对所述训练数据进行模型训练，生成预测模型。

6、根据权利要求 3-5 中任一项所述的在线预测模型的发布方法，其中，所述离线预测模型包括用户离线预测模型以及条目离线预测模型，所述获取离线预测模型针对多个样本用户标识生成的中间用户特征向量，以及针对多个样本条目标识生成的中间条目特征向量包括：

分别将每个样本用户标识及对应的用户属性信息输入至用户离线预测模型，并获取所述用户离线预测模型输出的每个样本用户标识对应的中间用户特征向量；

分别将每个样本条目标识及对应的条目属性信息输入至条目离线预测模型，并获取所述条目离线预测模型输出的每个样本条目标识对应的中间条目特征向量。

7、一种推荐装置，包括：

输入信息确定模块，设置为确定输入至已训练的在线预测模型中的目标用户标识以及多个候选条目标识，其中，所述在线预测模型中存储有多个样本用户标识对应的中间用户特征向量，以及多个样本条目标识对应的中间条目特征向量；

中间用户特征向量查找模块，设置为在所述在线预测模型的多个样本用户

标识中查找所述目标用户标识，获得所述目标用户标识对应的中间用户特征向量；

中间条目特征向量查找模块，设置为在所述在线预测模型的多个样本条目标识中查找每个候选条目标识，获得每个候选条目标识对应的中间条目特征向量；

预测处理模块，设置为在所述在线预测模型中，对所述目标用户标识对应的中间用户特征向量以及每个候选条目标识对应的中间条目特征向量进行预测处理，输出每个候选条目标识与所述目标用户标识的匹配度；

目标条目标识确定模块，设置为根据每个候选条目标识与所述目标用户标识的匹配度，从所述候选条目标识中确定用于推荐至所述目标用户标识的目标条目标识。

8、一种在线预测模型的发布装置，包括：

模型参数获取模块，设置为获取在线预测模型的模型参数和离线预设模型的模型参数；

中间向量获取模块，设置为获取离线预测模型针对多个样本用户标识生成的中间用户特征向量，以及针对多个样本条目标识生成的中间条目特征向量；

存储模块，设置为将每个样本用户标识对应的中间用户特征向量、每个样本条目标识对应的中间条目特征向量与所述模型参数存储至预设数据库中；

模型发布模块，设置为基于所述预设数据库对所述在线预测模型进行发布。

9、一种计算机设备，包括存储器、处理器及存储在存储器上并可在处理器上运行的计算机程序，所述处理器执行所述程序时实现以下至少之一：

如权利要求 1-2 中任一所述的推荐方法；如权利要求 3-6 中任一所述的在线模型的发布方法。

10、一种计算机可读存储介质，其上存储有计算机程序，所述程序被处理器执行时实现以下至少之一：

如权利要求 1-2 中任一所述的推荐方法；如权利要求 3-6 中任一所述的在线模型的发布方法。