

一种信息推荐的方法、装置、推荐服务器和存储介质

本申请要求在2019年10月31日提交中国专利局、申请号为201911053819.3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

本申请实施例涉及机器学习技术领域，例如一种信息推荐的方法、装置、推荐服务器和存储介质。

背景技术

随着人工智能交互技术的快速发展，预先训练的机器学习模型越来越广泛的应用在推荐、搜索、广告等各类业务的大数据场景下，针对用户的搜索需求，后台会获取大量相关的推荐信息，此时需要预测多条推荐信息的点击通过率（Click Through Rate, CTR），以便根据多个推荐信息的点击通过率为用户推送最优的信息。

目前通常采用逻辑回归算法训练得到尽量拟合用户对推荐信息的点击行为的业务预测模型，该业务预测模型中包含推荐信息中每一特征维度下的模型权重系数，在预测每一推荐信息的点击通过率时，需要去查询推荐信息中每一特征维度下的特征值以及对应的模型权重系数，此时由于推荐信息中存在大量高维稀疏特征，如推荐信息中身份维度下的用户身份标识（identity, ID）、主播ID或者设备ID等，最高能够达到上亿维，使得业务预测模型在存储推荐信息中的大量高维稀疏特征时会占用大量的内存空间，导致内存不足的情况；而且对于包含大量特征的推荐信息，需要耗费较长时间去查询推荐信息中每一特征维度下的模型权重系数，这使得业务预测模型任务量过大，导致推荐信息的预测效率较低。

发明内容

本申请实施例提供了一种信息推荐的方法、装置、推荐服务器和存储介质，减小了推荐模型的大小，在保证推荐准确性的同时，提高信息推荐效率。

第一方面，本申请实施例提供了一种信息推荐的方法，包括：

通过推荐模型对推荐信息的特征向量进行预测，得到所述推荐信息的点击通过率，其中，所述推荐模型中去掉了与所述特征向量的特征相同且频次小于或等于第一阈值的特征对应的模型参数；

根据所述点击通过率，下发所述推荐信息。

第二方面，本申请实施例提供了一种信息推荐的装置，该装置包括：

点击预测模块，设置为通过推荐模型对推荐信息的特征向量进行预测，得到所述推荐信息的点击通过率，其中，所述推荐模型中去掉了与所述特征向量的特征相同且频次小于或等于第一阈值的特征对应的模型参数；

信息推荐模块，设置为根据所述点击通过率，下发所述推荐信息。

第三方面，本申请实施例提供了一种推荐服务器，该推荐服务器包括：

至少一个处理器；

存储装置，设置为存储至少一个程序；

当所述至少一个程序被所述至少一个处理器执行，使得所述至少一个处理器实现本申请任意实施例所述的信息推荐的方法。

第四方面，本申请实施例提供了一种模型训练系统，该系统包括：训练服务器集群和参数服务器集群，所述参数服务器集群中存储预设的机器学习模型中不同特征对应的模型参数；

所述训练服务器集群依次从所述参数服务器集群获取所述机器学习模型中不同特征对应的最新模型参数，根据该模型参数对历史推荐信息的历史特征向量进行处理，得到对应的点击通过率，并计算所述机器学习模型中频次大于第一阈值的特征对应的模型参数的损失系数，将所述损失系数返回给所述参数服务器集群；

所述参数服务器集群根据该损失系数更新对应的模型参数并反馈给所述训

练服务器集群，直至所述机器学习模型完成训练后，去掉频次小于或等于第一阈值的特征对应的模型参数，得到剩余模型参数，并利用所述剩余模型参数生成对应的推荐模型。

第五方面，本申请实施例提供了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计算机程序，所述计算机程序被处理器执行时，实现本申请任意实施例所述的信息推荐的方法。

附图说明

图 1A 为本申请一实施例提供的一种信息推荐的方法的流程图；

图 1B 为本申请一实施例提供的信息推荐过程的原理示意图；

图 2A 为本申请一实施例提供的信息推荐的方法中训练推荐模型的流程图；

图 2B 为本申请一实施例提供的推荐模型训练过程的原理示意图；

图 3A 为本申请一实施例提供的另一种信息推荐的方法中训练推荐模型的流程图；

图 3B 为本申请一实施例提供的另一种信息推荐的方法中推荐模型的训练过程的原理示意图；

图 4 为本申请一实施例提供的又一种信息推荐的方法的流程图；

图 5 为本申请一实施例提供的一种信息推荐的装置的结构示意图；

图 6 为本申请一实施例提供的一种模型训练系统的原理架构图；

图 7 为本申请一实施例提供的一种推荐服务器的结构示意图。

具体实施方式

图 1A 为本申请一实施例提供的一种信息推荐的方法的流程图，本实施例可应用于任一种可以为用户提供针对相关推荐信息的下发请求服务的后台推荐服务器中。本申请实施例的技术方案适用于为用户推荐相关信息的情况中。本实施例提供的一种信息推荐的方法可以由本申请实施例提供的信息推荐的装置来

执行，该装置可以通过软件和/或硬件的方式来实现，并集成在执行本方法的推荐服务器中。

参考图 1A，该方法可以包括步骤 S110 至步骤 S120。

在步骤 S110 中，通过推荐模型对推荐信息的特征向量进行预测，得到推荐信息的点击通过率。

其中，推送信息是推送服务器根据接收到的用户在客户端作用的下发请求，根据用户的历史行为数据通过相关技术中的物品推荐 itemCF 算法、用户推荐 userCF 算法、内容推荐 content-based 算法或服务推荐 vector-based 算法等查找出的大量与本次下发请求对应的推荐信息。

示例性的，该下发请求为用户在需要推荐服务端在当前推荐对应的相关信息时，通过在客户端执行对应的内容刷新操作或者对具体某一信息的搜索查找操作，而对应生成本次请求中的指令信息。在下发请求中不包含具体的请求内容时，可以通过对该用户的历史行为数据进行分析，判断该用户较为感兴趣的新闻、资讯、音乐或短视频等内容，从而查找出对应的数据内容，作为本次请求对应的推荐信息，该种请求方法一般针对的是用户在客户端安装的对应软件产品中执行内容刷新操作，不确定所需要的具体信息，此时根据历史行为数据分析对应的推荐信息，从而为用户推荐优质的信息；同时，在下发请求中包含具体的请求内容时，可以直接对该具体的请求内容进行分析，从而查找与本次请求的具体内容相关联的新闻、资讯、音乐或短视频等内容，将查找到的关联信息作为本次请求对应的推荐信息，该种请求方法一般针对的是用户在客户端安装的对应软件产品中执行相关内容信息的搜索查找操作，此时可以根据本次请求中的具体搜索内容确定对应的推送信息，从而为用户推荐匹配的信息。

同时，推荐信息的特征向量是指根据推荐信息的各类特征分析出的在不同维度上的特征数据，例如，如果从点赞、转发和评论三方面对某一视频进行特征分析，则从这三方面分别对应分析出的点赞结果、转发结果和评论结果可以

组成该视频的一个三维特征向量。

此外，推荐模型为预先训练好的一种用于快速为用户筛选出最匹配的推荐信息的网络模型，该预先训练为针对每个历史推荐信息对应的特征向量，对用户每一个历史推荐信息的点击行为进行的训练，相关技术中的训练会不断更新网络模型中预设的与推荐信息对应的特征向量中每一特征相同的特征所对应的模型参数，因此所训练出的网络模型中存在与推荐信息对应的特征向量中每一特征对应的模型参数，以便对特征向量中的每个特征进行分析，此时由于在搜索、个性化推送或者广告通知等推荐场景下，每一推荐信息中包含大量不同类型下的特征，使得训练后的网络模型中也存在大量不同类型特征对应的模型参数，导致内存占用过大，而且在预测用户对推荐信息的点击行为时，需要耗费较长时间在大量模型参数中查询每一特征对应的模型参数，降低了模型运算的效率。

本实施例中考虑到推荐信息的特征向量中可能包含大量高维稀疏特征，如用户 ID、主播 ID 或者设备 ID 等，此时如果该类高维稀疏特征极少参与到用户对推荐信息的点击行为的训练过程中，从而对用户对不同推荐信息的点击行为不造成影响，因此在模型训练过程中，如图 1B 所示，本实施例可以排除对历史推荐信息的特征向量中包含的该类特征的训练，仅保留对用户的点击行为存在较大影响的特征的训练，使得训练后的推荐模型中仅保留对用户的点击行为存在较大影响的特征对应的模型参数。本实施例中可以在模型训练过程中根据历史推荐信息的特征向量，判断网络模型中预设的每一特征出现的次数（本实施例中特征的频次），该出现的次数表示该特征对用户的点击行为的影响程度，因此本实施例训练好的推荐模型中去掉了与特征向量的特征相同且频次小于或等于第一阈值的特征对应的模型参数，仅保留频次大于第一阈值的特征对应的模型参数，从而减少了推荐模型的大小，降低了推荐模型中的模型参数占用空间。示例性的，本实施例中推荐模型中的模型参数可以为该推荐模型中不同特征对应的权重系数，该权重系数用于表示每个特征对于用户的点击行为的影响

程度。

同时，推荐信息的点击通过率是指将该推荐信息从全部的推荐信息中筛选出来并推荐给用户所在客户端，并被用户点击查看的点击达到率，也就是该推荐信息的实际点击次数除以全部推荐信息的信息总量，以此确定下发给用户的每个推荐信息的优劣性。

示例性的，用户通过在客户端执行对应的内容刷新操作或者对具体某一信息的搜索查找操作，生成对应的下发请求，并发送给推荐服务器，该推荐服务器根据用户的历史行为数据通过相关技术中的相似度算法查找出大量与本次下发请求对应的推荐信息，并根据每个推荐信息的标识信息从存储有上传信息的不同类型特征的特征服务端中获取对应的特征向量，进而将每个推荐信息的特征向量输入到本实施例预先训练好的推荐模型中，通过该推荐模型中保留的频次大于第一阈值的特征对应的模型参数对推荐信息的特征向量中对应的特征进行分析，从而预测得到推荐信息的点击通过率。

在步骤 S120 中，根据点击通过率，下发推荐信息。

示例性的，在得到每个推荐信息的点击通过率后，可以明确用户对每个推荐信息的喜好程度，进而根据用户的下发请求和点击通过率在全部的推荐信息中筛选出对应的推荐信息，并下发给用户。示例性的，在用户当前需要获取最优的一条信息的情况下，将点击通过率最高的推荐信息下发给用户，在用户需要一次性获取多条相关信息的情况下，将点击通过率靠前的多条推荐信息下发给用户。

本实施例提供的技术方案，在推荐模型的训练过程中，去掉与推荐信息的特征向量中的特征相同且频次小于或等于第一阈值的特征对应的模型参数，仅保留频次大于第一阈值的特征对应的模型参数，从而减小了推荐模型的大小，降低了推荐模型中的模型参数的占用空间，同时在通过该推荐模型中的模型参数对推荐信息的特征向量进行预测时，无需查询特征向量中的每一特征对应的

模型参数，降低了查询过程中的耗费时长，仅对推荐信息的特征向量中与推荐模型中保留的模型参数对应的特征进行分析，在保证推荐准确性的同时，提高了信息推荐效率。

图 2A 为本申请一实施例提供的信息推荐的方法中训练推荐模型的流程图，图 2B 为本申请一实施例提供的推荐模型训练过程的原理示意图。本实施例是在上述实施例的基础上进行细化。本实施例中推荐模型的训练可以在线下执行，线下的推荐模型的训练与线上采用该训练好的推荐模型预测每个推荐信息的点击通过率的步骤可以在不同的设备上实现，因此本实施例中可以通过预先架构的模型训练系统来训练该推荐模型，在训练完成后，将该推荐模型发布到推荐服务器中，用于线上执行本申请任意实施例提及的信息推荐的方法。本实施例中主要对于推荐模型的线下训练过程进行详细的解释说明。

如图 2A 所示，本实施例中可以包括步骤 S210 至步骤 S230。

在步骤 S210 中，根据历史推荐信息的历史特征向量在预设的机器学习模型中的点击通过率，依次更新机器学习模型中频次大于第一阈值的特征对应的模型参数。

其中，在训练推荐模型前，会预先构建一个机器学习模型，该机器学习模型中针对不同类型下的每一特征均预先设置对应的模型参数的初始值，后续通过大量不同类型下的历史推荐信息的点击通过率的预测误差，不断对该机器学习模型中每个特征对应的模型参数进行修正，直至能够准确预测任一类型下的历史推荐信息的点击通过率。

示例性的，获取大量不同类型下的历史推荐信息，用于准确分析用户对不同类型信息的点击行为，进而判断用户的喜好；将多个历史推荐信息的历史特征向量输入到预设的机器学习模型中，由该机器学习模型中预设的当前模型参数对历史特征向量中对应的特征进行分析，进而得到每个历史推荐信息的点击通过率，通过该历史推荐信息的历史标签和点击通过率进行比对，判断机器

学习模型中当前模型参数下的预测误差，同时根据历史推荐信息的特征向量中包含的特征，判断该机器学习模型中每个特征出现的次数，也就是本实施例中含特征的频次，并将该预测误差在机器学习模型中反向传播，用于对选取出的机器学习模型中频次大于第一阈值的特征对应的模型参数进行修正，以便后续提高预测结果的准确性。

示例性的，如图 2B 所示，本实施例中根据历史推荐信息的历史特征向量在预设的机器学习模型中的点击通过率，依次更新机器学习模型中频次大于第一阈值的特征对应的模型参数，可以包括：通过机器学习模型对当前历史推荐信息的历史特征向量进行预测，得到对应的当前点击通过率；采用预设损失函数计算当前点击通过率在频次大于第一阈值的特征对应的模型参数下的损失系数，并根据该损失系数修正机器学习模型中对应的模型参数，继续通过机器学习模型对下一历史推荐信息的历史特征向量进行预测，并修正机器学习模型中频次大于第一阈值的特征对应的模型参数，直至机器学习模型完成训练。

示例性的，将当前历史推荐信息的历史特征向量输入预设的机器学习模型中，通过当前的模型参数对当前历史推荐信息的历史特征向量中的对应特征进行分析，预测得到该当前历史推荐信息的当前点击通过率，并根据历史推荐信息的历史特征向量中包含的特征，分析机器学习模型中每个特征的频次，进而标记出频次大于第一阈值的特征对应的模型参数；同时采用预设损失函数确定当前点击通过率在该机器学习模型中频次大于第一阈值的特征对应的模型参数下的损失系数，也就是该模型参数距离达到能够准确分析出历史推荐信息的点击通过率的差距，进而根据该损失系数分别修正机器学习模型中对应的模型参数；在本次修正后继续获取下一历史推荐信息的历史特征向量，并输入到修正后的机器学习模型中，由该机器学习模型中修正后的模型参数（频次大于第一阈值的特征对应的模型参数）继续分析下一个历史推荐信息的历史特征向量中与该模型参数对应的特征，排除频次小于或等于第一阈值的特征，得到下一个历史推荐信息的下一点击通过率，并继续修正该机器学习模型中频次大于第一

阈值的特征对应的模型参数，依次循环，直至历史推荐信息的点击通过率的预测误差低于预设误差阈值，表示此时该机器学习模型完成训练。

在步骤 S220 中，在机器学习模型完成训练后，去掉频次小于或等于第一阈值的特征对应的模型参数，得到剩余模型参数，并利用剩余模型参数生成推荐模型。

其中，剩余模型参数为机器学习模型中频次大于第一阈值的特征对应的模型参数。

示例性的，由于机器学习模型中频次小于或等于第一阈值的特征对于预测用户的点击行为几乎未作出贡献，因此为了推荐模型的效率，本实施例中在机器学习模型完成训练后，可以直接根据模型训练过程中对频次大于第一阈值的特征对应的模型参数的标记情况，去掉该机器学习模型中频次小于或等于第一阈值的特征对应的模型参数，而不对推荐信息的点击通过率的预测结果造成影响，后续利用去掉频次小于或等于第一阈值的特征对应的模型参数后的剩余模型参数（机器学习模型中频次大于第一阈值的特征对应的模型参数）生成对应的推荐模型，使得该推荐模型中仅保留频次大于第一阈值的特征对应的模型参数，在保证推荐准确性的基础上，减小了推荐模型的大小，提高了推荐模型的效率。

此外，为了准确判断出机器学习模型中频次小于或等于第一阈值的特征对应的模型参数，本实施例在依次更新机器学习模型中频次大于第一阈值的特征对应的模型参数之后，还可以包括：将机器学习模型中频次小于或等于第一阈值的特征对应的模型参数设置为预设稀疏值，以便在机器学习模型完成训练后，去掉预设稀疏值，得到剩余模型参数。

示例性的，在模型训练过程中，在将每一个历史推荐信息的历史特征向量输入到机器学习模型中进行训练的过程中，可以根据历史特征向量中包含的特征确定该机器学习模型中预设的每个特征的频次，此时对于频次大于第一阈值

的特征对应的模型参数进行修正，对于频次小于或等于第一阈值的特征对应的模型参数，由于后续该特征对应的模型参数不再参与学习训练，因此直接将频次小于或等于第一阈值的特征对应的模型参数设置为预设稀疏值（本实施例中该预设稀疏值为 0），后续在机器学习模型完成训练后，直接去掉设置为预设稀疏值的模型参数，进而保留频次大于第一阈值的特征对应的模型参数。

示例性的，本实施例中在机器学习模型中配置改进的跟随正则化逻辑（Followed the Regularized Leader, FTRL）算法，通过该改进的 FTRL 算法对在模型训练过程中依次对频次大于第一阈值的特征对应的模型参数进行修正。相关技术中的 FTRL 算法如下：

$$\begin{cases} z = z + g - w \left(\sqrt{n + g^2} - \sqrt{n} \right) / \alpha \\ n = n + g^2 \\ w = \begin{cases} 0; |z| \leq \lambda_1 \\ - \left(z - \text{sgn}(z) \lambda_1 \right) / \left(\frac{\beta + \sqrt{n_i}}{\alpha} + \lambda_2 \right); \text{otherwise} \end{cases} \end{cases};$$

其中， g 为某一特征对应的模型参数的损失系数， z 和 n 为机器学习模型在训练过程中的中间变量， w 为特征对应的模型参数， α 、 β 、 λ_1 和 λ_2 为稀疏性训练的限定参数。

本实施例中在上述 FTRL 算法的基础上，对机器学习模型中特征的频次进行限定，改进的 FTRL 算法如下：

$$\begin{cases} f = f + 1 \\ z = \begin{cases} 0; f \leq \lambda_3 \\ z + g - w(\sqrt{n + g^2} - \sqrt{n})/\alpha; \text{otherwise} \end{cases} \\ n = \begin{cases} 0; f \leq \lambda_3 \\ n + g^2; \text{otherwise} \end{cases} \\ w = \begin{cases} 0; f \leq \lambda_3 \text{ or } |z| \leq \lambda_1 \\ -(z - \text{sgn}(z)\lambda_1) / \left(\frac{\beta + \sqrt{n_1}}{\alpha} + \lambda_2 \right); \text{otherwise} \end{cases} \end{cases};$$

其中， f 为机器学习模型中某一特征的频次， λ_3 为特征的频次限定的第一阈值；改进后的 FTRL 算法对机器学习模型中频次大于第一阈值的特征对应的模型参数依然进行常规的训练，而直接将频次小于或等于第一阈值的特征对应的模型参数设置为 0，不参与后续的机器学习过程，也就是在机器学习模型的训练过程中，根据历史推荐信息的历史特征向量，确定机器学习模型中预设的每个特征的频次，只有特征频次大于第一阈值后才会对该特征对应的模型参数进行机器学习，否则该特征对应的模型参数一直设置为 0；后续在机器学习模型完成训练后，利用不为 0 的模型参数生成对应的推荐模型，在保证推荐准确性的基础上，减少了推荐模型的大小，提高了推荐模型的效率。

在步骤 S230 中，线上发布推荐模型。

示例性的，在生成对应的推荐模型后，可以直接将该推荐模型在线上发布，也就是应用到执行本申请任意实施例提供的信息推荐的方法的推荐服务器中，以便在线上直接采用该推荐模型预测每个推荐信息的点击通过率，为用户推荐更感兴趣的内容，提高信息推荐的准确性和效率。

本实施例提供的技术方案，通过机器学习模型对历史推荐信息预测的点击通过率不断更新机器学习模型中频次大于第一阈值的特征对应的模型参数，并在机器学习模型完成训练后，去掉频次小于或等于第一阈值的特征对应的模型参数，并利用剩余模型参数生成对应的推荐模型，无需每次对频次小于或等于第一阈值的特征对应的模型参数进行修正，提高了模型训练的高效性，减少了推荐模

型的大小，同时降低了模型参数和高维稀疏特征的内存占用。

图 3A 为本申请一实施例提供的另一种信息推荐的方法中训练推荐模型的流程图，图 3B 为本申请一实施例提供的另一种信息推荐的方法中推荐模型的训练过程的原理示意图。本实施例是在上述实施例的基础上进行细化。由于在机器学习模型完成训练后，在模型规模达到内存上限的情况下，会造成模型预测的拥塞，因此需要对频次较低的特征对应的模型参数再次进行过滤，以保证推荐模型的高效性。本实施例主要对于推荐模型的二次优化过程进行详细的解释说明。

如图 3A 所示，该方法可以包括步骤 S310 至步骤 S340。

在步骤 S310 中，根据历史推荐信息的历史特征向量在预设的机器学习模型中的点击通过率，依次更新机器学习模型中频次大于第一阈值的特征对应的模型参数。

在步骤 S320 中，在机器学习模型完成训练后，去掉频次小于或等于第一阈值的特征对应的模型参数，得到剩余模型参数，在机器学习模型规模达到内存上限的情况下，在剩余模型参数中过滤频次小于或等于第二阈值的特征对应的模型参数，并利用过滤后的剩余模型参数生成推荐模型。

示例性的，在机器学习模型完成训练后，可以去掉该机器学习模型中频次小于或等于第一阈值的特征对应的模型参数，此时在该机器学习模型的规模还是达到内存上限的情况下，在去掉频次小于或等于第一阈值的特征对应的模型参数后的剩余模型参数中继续过滤出频次小于或等于第二阈值的特征对应的模型参数，此时该第二阈值大于第一阈值，释放相应的内存空间，并利用过滤后的剩余模型参数生成对应的推荐模型，进而减少推荐模型的大小。

在步骤 S330 中，在机器学习模型完成当前时段下的训练后，根据预设的频次衰减系数更新机器学习模型中不同特征的频次，并在去掉频次小于或等于第一阈值的特征对应的模型参数后的剩余模型参数中过滤频次小于或等于第三阈

值的特征对应的模型参数，利用过滤后的剩余模型参数生成当前时段下的推荐模型。

示例性的，由于在信息推荐过程中不断存在新的推荐业务，从而存在新的特征，而如果某一时段更新的推荐业务特征仅热门一段时间，后期不再对该业务特征进行参考，但机器学习模型中在热门时段已经记录该业务特征的频次大于第一阈值，在后期训练时还是会一直参考该业务特征，但该业务特征已经不再对模型训练有任何贡献，因此为了保证推荐模型的准确性，本实施例可以对机器学习模型进行分时段训练，获取当前时段下的历史推荐信息，不断对机器学习模型中频次大于第一阈值的特征对应的模型参数进行修正，在机器学习模型完成当前时段下的训练后，由于前一时段的某一特征在当前时段可能不再出现，因此根据预设的频次衰减系数更新机器学习模型中不同特征的频次，此时如果某一特征一直不出现，则该特征的频次会不断衰减，直至小于或等于第三阈值而被去掉，如图 3B 所示，本实施例中首先去掉频次小于或等于第一阈值的特征对应的模型参数，得到剩余模型参数，然后根据预设的频次衰减系数更新机器学习模型中不同特征的频次，之后在剩余模型参数中再次过滤频次小于或等于第三阈值的特征对应的模型参数，进而利用过滤后的剩余模型参数生成当前时段下的推荐模型，提高推荐模型的准确性和高效性。

需要说明的是，本实施例中过滤频次小于或等于第三阈值的特征对应的模型参数可以根据训练后的机器学习模型的规模来判断是否执行，在机器学习模型完成当前时段下的训练后规模未达到内存上限的情况下，可以不用再次过滤频次小于或等于第三阈值的特征对应的模型参数，而直接利用去掉频次小于或等于第一阈值的特征对应的模型参数后的剩余模型参数生成对应的推荐模型。本实施例中 S320 和 S330 可以选择执行其中的一个。

在步骤 S340 中，线上发布推荐模型。

本实施例提供的技术方案，在机器学习模型完成训练后，在机器学习模型

的规模达到内存上限的情况下，在去掉频次小于或等于第一阈值的特征对应的模型参数后的剩余模型参数中再次过滤频次小于或等于第二阈值的特征对应的模型参数，利用过滤后的剩余模型参数生成对应的推荐模型，减少了推荐模型的大小，降低了模型参数和高维稀疏特征的内存占用；同时通过分时段训练机器学习模型，并根据频次衰减系数不断更新不同特征的频次，保证不同时段下推荐模型的准确性和高效性。图 4 为本申请一实施例提供的又一种信息推荐的方法的流程图。本实施例是在上述实施例的基础上进行细化。本实施例中对于具体的信息推荐过程进行说明，如图 4 所示，本实施例中可以包括步骤 S410 至步骤 S430。

在步骤 S410 中，获取推荐模型中去掉与特征向量的特征相同且频次小于或等于第一阈值的特征对应的模型参数后的剩余模型参数。

示例性的，在根据用户请求获取到大量推荐信息后，需要通过线上发布的推荐模型对每个推荐信息的特征向量进行预测，因此首先确定线上发布的推荐模型，并查找出该推荐模型中去掉与特征向量的特征相同且频次小于或等于第一阈值的特征对应的模型参数后的剩余模型参数，也就是频次大于第一阈值的特征对应的模型参数，以便对每个推荐信息的特征向量中对应的特征进行分析。

在步骤 S420 中，根据剩余模型参数以及特征向量中与剩余模型参数对应的特征，计算推荐信息的点击通过率。

示例性的，根据推荐模型中保留的剩余模型参数，在每个推荐信息的特征向量中筛选出该剩余模型参数对应的特征，此时模型参数可以是特征的权重系数，采用加权求和的方式将每一剩余模型参数和推荐信息的特征向量中与该剩余模型参数对应的特征相乘并求和，将推荐信息的特征向量中每个剩余模型参数对应的特征加权求和的结果，作为本实施例中推荐信息的点击通过率，从而减少推荐模型中的计算量，提高信息推荐的效率。

在步骤 S430 中，根据点击通过率对推荐信息排序，并构建对应的下发列表

下发给用户。

示例性的，在得到多个推荐信息的点击通过率后，根据该点击通过率的高低为多个推荐信息进行排序，并根据用户的需求选择排序靠前的推荐信息构建对应的下发列表，并将该下发列表显示给用户，保证信息推荐的准确性。

本实施例提供的技术方案，在推荐模型的训练过程中，去掉与推荐信息的特征向量中的特征相同且频次小于或等于第一阈值的特征对应的模型参数，仅保留频次大于第一阈值的特征对应的模型参数，从而减小了推荐模型的大小，降低了推荐模型中的参数占用空间，同时在通过该推荐模型中的模型参数对推荐信息的特征向量进行预测时，无需查询特征向量中的每一特征对应的模型参数，降低了查询过程中的耗费时长，仅对推荐信息的特征向量中与推荐模型中保留的模型参数对应的特征进行分析，在保证推荐准确性的同时，提高了信息推荐效率，进而为用户推送更加优质的内容，改善了信息推送效果。

图 5 为本申请一实施例提供的一种信息推荐的装置的结构示意图，如图 5 所示，该装置可以包括：

点击预测模块 510，设置为通过推荐模型对推荐信息的特征向量进行预测，得到推荐信息的点击通过率，其中，该推荐模型中去掉了与特征向量的特征相同且频次小于或等于第一阈值的特征对应的模型参数；

信息推荐模块 520，设置为根据点击通过率，下发推荐信息。

本实施例提供的技术方案，在推荐模型的训练过程中，去掉与推荐信息的特征向量中的特征相同且频次小于或等于第一阈值的特征对应的模型参数，仅保留频次大于第一阈值的特征对应的模型参数，从而减小了推荐模型的大小，降低了推荐模型中的参数占用空间，同时在通过该推荐模型中的模型参数对推荐信息的特征向量进行预测时，无需查询特征向量中的每一特征对应的模型参数，降低了查询过程中的耗费时长，仅对推荐信息的特征向量中与推荐模型中保留的模型参数对应的特征进行分析，在保证推荐准确性的同时，提高了信息

推荐效率。

在一实施例中，上述点击预测模块 510，可以设置为：

获取推荐模型中去掉与特征向量的特征相同且频次小于或等于第一阈值的特征对应的模型参数后的剩余模型参数；

根据剩余模型参数以及特征向量中与剩余模型参数对应的特征，计算推荐信息的点击通过率。

在一实施例中，上述推荐模型可以通过模型训练系统包括的模型参数更新模块和推荐模型生成模块执行下述操作生成：

模型参数更新模块，设置为根据历史推荐信息的历史特征向量在预设的机器学习模型中的点击通过率，依次更新机器学习模型中频次大于第一阈值的特征对应的模型参数；

推荐模型生成模块，设置为在机器学习模型完成训练后，去掉频次小于或等于第一阈值的特征对应的模型参数，得到剩余模型参数，并利用最新的剩余模型参数生成推荐模型，该剩余模型参数为机器学习模型中频次大于第一阈值的特征对应的模型参数。

在一实施例中，上述模型参数更新模块，可以设置为：

通过机器学习模型对当前历史推荐信息的历史特征向量进行预测，得到对应的当前点击通过率；

采用预设损失函数计算当前点击通过率在频次大于第一阈值的特征对应的模型参数下的损失系数，并根据该损失系数修正机器学习模型中对应的模型参数，继续通过机器学习模型对下一个历史推荐信息的历史特征向量进行预测，并修正机器学习模型中频次大于第一阈值的特征对应的模型参数，直至机器学习模型完成训练。

在一实施例中，上述模型训练系统，还可以包括：

稀疏参数设置模块，设置为将机器学习模型中频次小于或等于第一阈值的特征对应的模型参数设置为预设稀疏值，以便在机器学习模型完成训练后，去掉设置为预设稀疏值的模型参数。

在一实施例中，上述推荐模型生成模块，可以设置为：

在机器学习模型规模达到内存上限的情况下，在剩余模型参数中过滤频次小于或等于第二阈值的特征对应的模型参数，并利用过滤后的剩余模型参数生成推荐模型，该第二阈值大于第一阈值。

在一实施例中的，上述推荐模型生成模块，还可以设置为：

在机器学习模型完成当前时段下的训练后，根据预设的频次衰减系数更新机器学习模型中不同特征的频次，并在去掉频次小于或等于第一阈值的特征对应的模型参数后的剩余模型参数中过滤频次小于或等于第二阈值的特征对应的模型参数，利用过滤后的剩余模型参数生成当前时段下的推荐模型。

在一实施例中的，上述模型训练系统，还可以包括：

模型发布模块，设置为线上发布推荐模型。

在一实施例中，上述模型参数为推荐模型中不同特征的权重系数。

在一实施例中，上述信息推荐模块 520，可以设置为：

根据点击通过率对推荐信息排序，并构建对应的下发列表下发给用户。

本实施例提供的信息推荐的装置可适用于上述任意实施例提供的信息推荐的方法。

图 6 为本申请一实施例提供的一种模型训练系统的原理架构图。本实施例是在上述实施例的基础上进行细化。参照图 6，该模型训练系统可以包括：训练服务器集群 610 和参数服务器集群 620，参数服务器集群 620 中存储预设的机器学习模型中不同特征对应的模型参数。

示例性的，训练服务器集群 610 依次从参数服务器集群 620 获取机器学习

模型中不同特征对应的最新模型参数，根据该模型参数对历史推荐信息的历史特征向量进行处理，得到对应的点击通过率，并计算机器学习模型中频次大于第一阈值的特征对应的模型参数的损失系数，将该损失系数返回给参数服务器集群 620；参数服务器集群 620 根据该损失系数更新对应的模型参数并反馈给训练服务器集群 610，直至机器学习模型完成训练后，去掉频次小于或等于第一阈值的特征对应的模型参数，得到剩余模型参数，并利用所述剩余模型参数生成对应的推荐模型。

同时，训练服务器集群 610 将机器学习模型中频次小于或等于第一阈值的特征对应的模型参数设置为预设稀疏值，并返回给参数服务器集群 620；参数服务器集群 620 将频次小于或等于第一阈值的特征对应的模型参数修正为预设稀疏值，并在机器学习模型完成训练后，去掉设置为预设稀疏值的模型参数。

在一实施例中，为了保证模型的高效性，参数服务器集群 620 在机器学习模型完成训练后，在所述机器学习模型规模达到内存上限的情况下，在剩余模型参数中过滤频次小于或等于第二阈值的特征对应的模型参数，并利用过滤后的剩余模型参数生成推荐模型，第二阈值大于第一阈值。

示例性的，为了保证每个时段下的推荐模型的准确性，参数服务器集群 620 在机器学习模型完成当前时段下的训练后，根据预设的频次衰减系数更新机器学习模型中不同特征的频次，并在去掉频次小于或等于第一阈值的特征对应的模型参数后的剩余模型参数中过滤频次小于或等于第三阈值的特征对应的模型参数，利用过滤后的剩余模型参数生成当前时段下的推荐模型。

此外，模型训练系统中还可以包括推荐服务器 630，参数服务器集群 620 将生成的推荐模型发布到推荐服务器 630 中，推荐服务器 630 设置为执行本申请任意实施例提供的信息推荐的方法。

需要说明的是，训练服务器集群 610 中分布式部署多个训练服务器，该训练服务器可以并行执行机器学习模型的训练，提高模型的训练速率；同时，参

数服务器集群 620 中分布式部署多个参数服务器，该参数服务器中分片存储机器学习模型中不同特征对应的模型参数，减少数据的存储压力；而且将模型训练过程中的训练步骤和参数修正步骤分离开来，降低了模型训练的计算压力。

本实施例提供的技术方案，通过训练服务器集群和参数服务器集群共同训练推荐模型，从而将训练步骤和参数更新步骤分离开来，将训练和模型参数分散到不同的计算节点上，减小单台服务器的计算压力，提高了推荐模型的训练速率。

图 7 为本申请一实施例提供的一种推荐服务器的结构示意图，如图 7 所示，该推荐服务器包括处理器 70、存储装置 71 和通信装置 72；推荐服务器中处理器 70 的数量可以是至少一个，图 7 中以一个处理器 70 为例；推荐服务器中的处理器 70、存储装置 71 和通信装置 72 可以通过总线或其他方式连接，图 7 中以通过总线连接为例。

存储装置 71 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，如本申请任意实施例所述的信息推荐的方法对应的程序指令/模块。处理器 70 通过运行存储在存储装置 71 中的软件程序、指令以及模块，从而执行推荐服务器的各种功能应用以及数据处理，即实现上述信息推荐的方法。

存储装置 71 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，存储装置 71 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，存储装置 71 可包括相对于处理器 70 远程设置的存储器，这些远程存储器可以通过网络连接至推荐服务器。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。

通信装置 72 可设置为实现推荐服务器与客户端之间的网络连接或者移动数据连接。

本实施例提供了一种推荐服务器可用于执行上述任意实施例提供的信息推荐的方法。

本申请一实施例还提供了一种计算机可读存储介质，计算机可读存储介质存储有计算机程序，该计算机程序被处理器执行时，可实现上述任意实施例中的信息推荐的方法。该方法可以包括：

通过推荐模型对推荐信息的特征向量进行预测，得到推荐信息的点击通过率，其中，该推荐模型中去掉了与特征向量的特征相同且频次小于或等于第一阈值的特征对应的模型参数；

根据点击通过率，下发推荐信息。

当然，本申请实施例所提供的一种包含计算机可执行指令的存储介质，其计算机可执行指令不限于如上所述的方法操作，还可以执行本申请任意实施例所提供的信息推荐的方法中的相关操作。

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，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，包括多个指令用以使得一台计算机设备（可以是个人计算机，服务器，或者网络设备等等）执行本申请各个实施例所述的方法。

值得注意的是，上述信息推荐的装置的实施例中，所包括的各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，各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。