(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 2 月 9 日 (09.02.2023)

WIPO | PCT

(10) 国际公布号
WO 2023/011382 A1

(51) 国际专利分类号:
G06F 16/9535 (2019.01)

(21) 国际申请号: PCT/CN2022/109297

(22) 国际申请日: 2022 年 7 月 30 日 (30.07.2022)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
202110877429.9 2021年7月31日 (31.07.2021) CN

(71) 申请人: 华为技术有限公司 (**HUAWEI TECHNOLOGIES CO., LTD.**) [CN/CN]; 中国广东省深圳市龙岗区坂田华为总部办公楼, Guangdong 518129 (CN)。

(72) 发明人: **郭威(GUO, Wei)**; 中国广东省深圳市龙岗区坂田华为总部办公楼, Guangdong 518129 (CN)。 **秦佳锐(QIN, Jiarui)**; 中国广东省深圳市龙岗区坂田华为总部办公楼, Guangdong 518129 (CN)。 **唐睿明(TANG, Ruiming)**; 中国广东省深圳市龙岗区坂田华为总部办公楼, Guangdong 518129 (CN)。 **刘志容(LIU, Zhirong)**; 中国广东省深圳市龙岗区坂田华为总部办公楼, Guangdong 518129 (CN)。 **何秀强(HE, Xiuqiang)**; 中国广东省深圳市龙岗区坂田华为总部办公楼, Guangdong 518129 (CN)。 **张伟楠(ZHANG, Weinan)**; 中国上海市闵行区东川路800号上海交通大学, Shanghai 200240 (CN)。 **俞勇(YU, Yong)**; 中国上海市闵行区东川路800号上海交通大学, Shanghai 200240 (CN)。

(74) 代理人: 广州三环专利商标代理有限公司 (**SCIHEAD IP LAW FIRM**); 中国广东省广州市越秀区先烈中路 80 号汇华商贸大厦 1508 室, Guangdong 510070 (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,

(54) **Title:** RECOMMENDATION METHOD, RECOMMENDATION MODEL TRAINING METHOD, AND RELATED PRODUCT

(54) **发明名称**: 推荐方法、推荐模型训练方法及相关产品

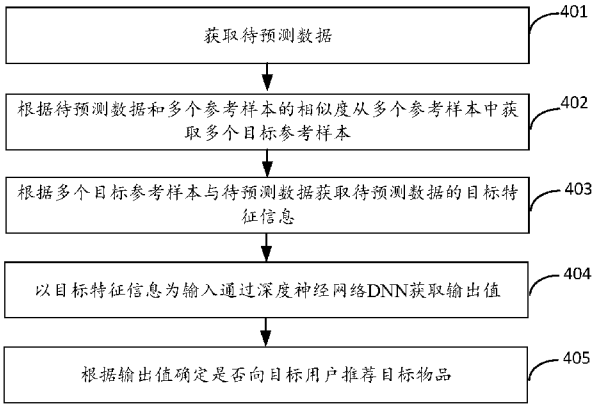


图 4

401 Obtain data to be predicted
402 Obtain a plurality of target reference samples from a plurality of reference samples according to the similarity between the data and the plurality of reference samples
403 Obtain target feature information of the data according to the plurality of target reference samples and the data
404 Obtain an output value by using the target feature information as an input by means of a deep neural network (DNN)
405 Determine, according to the output value, whether to recommend a target article to the target user

(57) **Abstract:** Embodiments of the present application provide a recommendation method, a recommendation model training method, and a related product. The recommendation method comprises: obtaining data to be predicted; obtaining a plurality of target reference samples from a plurality of reference samples according to the similarity between the data and the plurality of reference samples, each reference sample and the data both comprising user feature domain data and article feature domain data, the user feature domain data of the data being used for indicating a target user feature, the article feature domain data of the data being used for indicating a target article feature, and each target reference sample and the data having the user feature domain data and/or the article feature domain data which are partially identical; obtaining target feature information of the data according to the plurality of target reference samples and the data; obtaining an output value by using the target feature information as an input by means of a deep neural network (DNN); and determining, according to the output value, whether to recommend a target article to the target user. The embodiments of the present application improve the recommendation accuracy.

WO 2023/011382 A1

[见续页]

BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CV, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IQ, 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, 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：本申请实施例提供了一种推荐方法、推荐模型训练方法及相关产品。该推荐方法包括：获取待预测数据；根据待预测数据和多个参考样本的相似度从多个参考样本中获取多个目标参考样本；每个参考样本和待预测数据均包括用户特征域数据和物品特征域数据，待预测数据的用户特征域数据用于指示目标用户特征，待预测数据的物品特征域数据用于指示目标物品特征，每个目标参考样本和待预测数据具有部分相同的用户特征域数据和/或物品特征域数据；根据多个目标参考样本与待预测数据获取待预测数据的目标特征信息；以目标特征信息为输入通过深度神经网络DNN获取输出值；根据输出值确定是否向目标用户推荐目标物品。本申请实施例有利于提高推荐精度。

推荐方法、推荐模型训练方法及相关产品

本申请要求于2021年07月31日提交中国专利局、申请号为202110877429.9、申请名称为“推荐方法、推荐模型训练方法及相关产品”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

本发明涉及人工智能技术领域，具体涉及一种推荐方法、推荐模型训练方法及相关产品。

背景技术

基于表格数据预测用户的行为是一项重要的任务，有着大量的真实应用，比如，在线广告中的点击率（Click-Through-Rate，CTR）预测，推荐系统中的评分预测以及商品排序，诈骗账户检测等等。这些应用场景收集到的数据都以表格式的格式进行存储。表格中的每一行对应一个样本，每个样本的每一列对应一个独特的特征。

从表格式数据中提取有价值的关系或者模式是机器学习系统学习准确的关键。为了更好的利用表格式数据，如何充分挖掘表格式数据行之间以及列之间蕴含的丰富信息就变得至关重要。早期的模型，如逻辑回归，支持向量机以及树模型将行样本作为输入来做预测。深度模型将行样本的类别特征映射成嵌入向量，然后使用单个行样本的特征向量进行用户行为预测。近年来，基于特征交互的模型和基于用户序列的模型成为了表格式数据建模的主流。基于特征交互的模型，致力于对表格式数据中每个行样本的列与列之间的特征进行交互，以充分挖掘用户序列特征去预测用户行为，并基于预测出的用户行为进行推荐。

然而，上述都是孤立的使用单个样本预测用户行为，导致基于预测出的用户行为进行推荐的精度较低。

发明内容

本申请提供了一种推荐方法、推荐模型训练方法及相关产品，通过融合目标参考样本的特征信息进行推荐，提高推荐精度。

第一方面，本申请实施例提供一种推荐方法，包括：获取待预测数据；根据待预测数据和多个参考样本的相似度从多个参考样本中获取多个目标参考样本；每个参考样本和待预测数据均包括用户特征域数据和物品特征域数据，待预测数据的用户特征域数据用于指示目标用户特征，待预测数据的物品特征域数据用于指示目标物品特征，每个目标参考样本和待预测数据具有部分相同的用户特征域数据和/或物品特征域数据；根据多个目标参考样本与待预测数据获取待预测数据的目标特征信息；目标特征信息包括第一目标特征向量组和第二目标特征向量组，第一目标特征向量组为向量化后的待预测数据，第二目标特征向量组为对多个目标参考样本进行向量化后融合得到；以目标特征信息为输入通过深度神经网络 DNN 获取输出值；根据输出值确定是否向目标用户推荐目标物品。

其中，输出值可以是一个概率值，该概率值反映了目标用户对目标物品进行操作的概率。对于不同的目标物品，目标用户对目标物品进行操作的概率有不同的理解。比如，当目标物

品为应用程序时，目标用户对目标物品进行操作的概率可以理解为目标用户点击应用程序的概率；再如，当目标物品为歌曲时，目标用户对目标物品进行操作的概率可以理解为目标用户喜欢该歌曲的概率；又如，当目标物品为商品时，目标用户对目标物品进行操作的概率可以理解为目标用户购买该商品的概率。

实际应用中，获取了概率值之后，可以对概率值进行后处理，得到输出值。比如，当概率值大于概率阈值时，以 1 作为输出值；当概率值小于或等于阈值时，则以 0 作为输出值，其中，0 表示目标用户不会操作目标物品，1 表示目标用户会操作目标物品。

针对单物品推荐场景来说，当输出值大于阈值时，确定向目标用户推荐目标物品，否则确定不向目标用户推荐该目标物品。此外，当本申请的方案应用到从多个候选物品中选出物品进行推荐的场景时，可以获取每个候选物品对应的输出值；然后，向目标用户推荐输出值最高的候选物品，或者，对多个候选物品的输出值进行排序，将排序靠前的（比如，前十个候选物品）候选物品推荐给目标用户。举例来说，进行歌曲推荐时，可以获取歌曲库中每个候选歌曲的输出值，然后，将输出值排名前十的歌曲推荐给目标用户。

可以看出，本申请中所获取的目标特征信息中除了包含待预测数据本身的特征信息之外，还包含有多个目标参考样本向量化融合之后的特征信息。由于目标参考样本是通过待预测数据与多个参考样本之间的相似度从多个参考样本中选取的，且与待预测数据具有部分相同的用户特征域数据和/或物品特征域数据，因此目标参考样本是多个参考样本中与待预测数据比较相似的参考样本，故目标参考样本中的用户行为可以为目标用户的行为的预测提供参考和经验，从而在使用融合目标参考样本的特征的目标特征信息进行输出值的预测时，可以使预测出的输出值比较精确，基于这个输出值进行物品推荐，提高了推荐的精度。

在一些可能的实施方式中，多个目标参考样本还包括标签数据；第二目标特征向量组为对多个目标参考样本进行向量化后融合得到，具体为：第二目标特征向量组为对多个目标参考样本的用户特征域数据，物品特征域数据以及标签数据进行向量化后融合得到。

其中，目标参考样本的用户特征数据用于表示参考用户特征，目标参考样本的物品特征数据用于表示参考物品特征。由于目标参考样本还携带有标签数据，即参考用户对参考物品的真实操作行为。因此，第二特征向量组包含有参考用户对参考物品的真实操作行为，则在使用目标特征信息对目标用户的行为进行预测时，可以结合参考用户对参考物品的真实操作行为预测目标用户对目标物品的操作行为，得到输出值，使预测出的输出值精度较高，进而提高物品推荐精度。

在一些可能的实施方式中，目标特征信息还包括第三目标特征向量组，第三目标特征向量组为对第一向量组中的目标特征向量进行两两交互得到，第一向量组包括第一目标特征向量组和第二目标特征向量组。

应说明，上述是对第一向量组中的目标特征向量进行两两交互，但在实际应用可以自由的进行两两交互。示例性的，可以对第一目标特征向量组中的多个第一目标特征向量进行两两交互，得到多个第三目标特征向量；或者，对第二目标特征向量组中的多个第二目标特征向量进行两两交互，得到多个第三目标特征向量，等等。

可以看出，在本实施方式中，对第一向量组中的目标特征向量进行两两交互，得到多个第三目标特征向量，从而使目标特征信息中还包含有高阶特征信息，即第三目标特征向量可以表征用户各个行为之间的联系，因此使用更高阶的特征信息进行行为预测，可进一步提高输出值的精度。举例来说，某个第一目标特征向量表示用户的年龄为 28 岁时，另外一个第一目标特征向量表示用户为男性时，则对这两个目标特征向量进行交互后得到的第三目标特征

向量表示用户是一个28岁的男性。单独使用每个目标特征向量进行预测时，则目标物品满足28岁的人的需求或者满足男性的需求，则认为目标用户有一定的概率去操作这个目标物品，得到的输出值一般会大于概率阈值，而将目标特征向量交互后，则只有当目标物品满足28岁的男性需求时，目标用户才有一定的概率去操作这个目标物品，得到的输出值才会大于概率阈值，因此，使得到的输出值的精度比较高，进一步提高推荐精度。

在一些可能的实施方式中，对第一目标特征向量组中的多个第一目标特征向量进行拼接，得到待预测数据的第二特征向量；对每个目标参考样本的多个第一特征向量进行拼接，得到每个目标参考样本的第二特征向量，每个目标参考样本的多个第一特征向量为对目标参考样本进行向量化得到；获取每个目标参考样本的第二特征向量与待预测数据的第二特征向量之间的相似度；根据每个目标参考样本的第二特征向量与待预测数据的第二特征向量之间的相似度，确定每个目标参考样本的权重；根据每个目标参考样本的权重，对多个目标参考样本在同一个特征域下的第一特征向量进行融合，得到第二目标特征向量组。

可以看出，在本实施方式中，通过注意力机制，从而使多个目标参考样本中与待预测数据关联程度最高的目标参考样本的权重最大，这样融合得到的第二目标特征向量主要指示的特征信息为关联程度最高的目标参考样本的特征信息，尽可能的多使用关联程度最高的目标参考样本指引对目标用户行为的预测，使预测出的目标用户对目标物品进行操作的概率精度更高，进而提高物品推荐的精度。

在一些可能的实施方式中，根据所述待预测数据和多个参考样本的相似度从所述多个参考样本中获取多个目标参考样本之前，方法还包括：获取多个原始样本，其中，每个所述原始样本包括用户特征域数据和物品特征域数据；将所述待预测数据的多个所述用户特征域数据和多个所述物品特征域数据作为元素，对所述多个原始样本进行倒排索引，得到所述多个参考样本。

可选的，首先对多个原始样本进行倒排，得到倒排列表；比如，可以将每个参考样本的用户特征域数据和物品特征域数据作为元素进行倒排，得到如表2所示的倒排列表，比如，倒排列表中的每一行中的第一列为元素，即该多个参考样本下的一个域数据（用户特征域数据或者物品特征域数据），第二列为多个参考样本中包含有该域数据的参考样本。有了倒排列表之后，将待预测数据的每个用户特征域数据和每个物品特征域数据作为元素，对多个原始样本进行索引，得到多个参考样本。即根据倒排列表中的对应关系，可索引出与每个用户特征域数据对应的参考样本，以及与每个物品特征域对应的参考样本；然后，将每个用户特征域数据对应的参考样本，以及与每个物品特征域对应的参考样本进行合并与去重，得到该多个参考样本。例如，待预测数据为[U4、LA、Student、L2、cell phone、B3]，将U4、LA、Student、L2、cell phone、B3均作为查询词，从表2所示的倒排列表中获取与LA对应的参考样本为[样本1、样本3]，与Student对应的参考样本为[样本1、样本2、样本3]，与L2对应的参考样本为[样本3]，与cell phone对应的参考样本为[样本3、样本4]，与B3对应的参考样本为[样本4]。然后，对从倒排列表中获取到的所有参考样本进行合并与去重，得到多个参考样本，即[样本1、样本2、样本3、样本4]。

可以看出，在本实施方式中，通过倒排对多个原始样本进行排序，得到倒排列表。由于使用了倒排列表，从而可以使用倒排列表从多个原始样本中快速索引出多个参考样本，排除掉一部分不相关的原始样本，这样就可以不用和每个原始样本进行相似度的计算，减轻计算压力，可以快速的筛选出目标参考样本，提高物品推荐的效率。

第二方面，本申请实施例提供一种推荐模型训练方法，推荐模型包括特征信息提取网络

和深度神经网络DNN，方法包括：获取多个训练样本，其中，每个训练样本包括用户特征域数据和物品特征域数据；根据第一训练样本和多个第二训练样本的相似度从多个第二训练样本中获取多个目标训练样本，其中，第一训练样本为多个训练样本中的一个，多个第二训练样本为多个训练样本除第一训练样本之外的部分或全部，第一训练样本的用户特征域数据用于指示第一参考用户特征，第一训练样本的物品特征域数据用于指示第一参考物品特征，第一训练样本和每个目标训练样本具有部分相同的用户特征域数据和/或物品特征域数据；将第一训练样本和多个目标训练样本输入到特征信息提取网络，得到第一训练样本的目标特征信息，其中，目标特征信息包括第四目标特征向量组和第五目标特征向量组，第四目标特征向量组为通过特征信息提取网络对第一训练样本进行向量化得到，第五目标特征向量组为通过特征信息提取网络对多个目标训练样本进行向量化后融合得到；将目标特征信息输入到深度神经网络DNN，得到输出值，输出值用于表征第一参考用户对第一参考物品进行操作的概率；根据输出值以及第一训练样本的标签数据进行推荐模型的训练，获得目标推荐模型。

应说明，将第一训练样本以及多个目标训练样本输入到推荐模型的特征信息提取网络，构造信息更丰富的目标特征信息，使目标特征信息中既包含有第一训练样本的特征信息，即多个第四目标特征向量，还包含有多个目标训练样本向量化后融合的特征信息，即多个第五目标特征向量，并且目标训练样本是通过第一训练样本与多个第二训练样本之间的相似度从多个第二训练样本中选取的，因此目标训练样本是与第一训练样本比较相似的训练样本，从而在使用第一训练样本的目标特征信息进行模型训练时，可以参考多个目标训练样本向量化后融合的特征信息（即先验知识）进行用户行为预测，得到输出值，使预测出的输出值更加精确，从而使训练过程中得到的损失比较小，模型更容易收敛；此外，由于参考了多个目标训练样本的用户特征信息，使模型能够记住更丰富的用户特征信息，从而使训练出模型精度较高，鲁棒性比较强。

在一些可能的实施方式中，第五目标特征向量组为对多个目标训练样本进行向量化后融合得到，具体为：第五目标特征向量组为通过特征信息提取网络对多个目标训练样本的用户特征域数据、物品特征域数据以及标签数据进行向量化后融合得到。

可以看出，在本实施方式中，目标训练样本携带有标签数据，由于每个目标训练样本的标签数据反映了每个目标训练样本中的用户对物品的真实操作行为。因此，使用目标特征信息对目标用户的行为进行预测时，可以结合目标训练样本中的用户对物品的真实操作行为，去预测第一训练样本中的第一参考用户对第一参考物品进行操作的概率，从而使预测出的输出值的精度较高，由于预测的输出值精度较高，使训练过程中得到的损失比较小，缩短模型训练周期，提高模型收敛速度。

在一些可能的实施方式中，目标特征信息还包括第六目标特征向量组，第六目标特征向量组是通过特征信息提取网络对第二向量组中的目标特征向量进行两两交互得到，第二向量组包括第四目标特征向量组和第五目标特征向量组。

可以看出，在本实施方式中，对第二向量组中的目标特征向量进行两两交互，得到多个第六目标特征向量，从而使目标特征信息中还包含有高阶特征信息，即第六目标特征向量可以表征第一参考用户的高阶特征，因此使用高阶特征进行行为预测，可进一步提高对用户行为的预测精度，进一步提高模型收敛速度。举例来说，某个第四目标特征向量表示用户的年龄为28岁时，另外一个第四目标特征向量表示用户为男性时，则对这两个第四目标特征向量进行交互后得到的第六目标特征向量表示用户是一个28岁的男性。单独使用每个第四目标特征向量进行预测时，则物品满足28岁的人的需求或者满足男性的需求，则认为用户有一定的

概率去操作这个物品，而将目标特征向量交互后，则只有当物品满足28岁的男性需求时，用户才有一定的概率去操作这个物品，从而提高了对用户行为的预测精度。

在一些可能的实施方式中，融合包括：对第四目标特征向量组中的多个第四目标特征向量进行拼接，得到第一训练样本的第二特征向量；对每个目标训练样本的多个第一特征向量进行拼接，得到每个目标训练样本的第二特征向量，每个目标训练样本的多个第一特征向量为对目标训练样本进行向量化得到；获取每个目标训练样本的第二特征向量与第一训练样本的第二特征向量之间的相似度；根据每个目标训练样本的第二特征向量与第一训练样本的第二特征向量之间的相似度，确定每个目标训练样本的权重；根据每个目标训练样本的权重，对多个目标训练样本在同一个特征域下的第一特征向量进行融合，得到第五目标特征向量组。

可以看出，通过注意力机制，从而使多个目标训练样本中与第一训练样本关联程度最高的目标训练样本的权重最大，这样融合得到的第五目标特征向量主要指示的特征信息为该目标训练样本的特征信息，从而尽可能的多使用关联程度最高的目标训练样本指引对第一参考用户行为的预测，从而使预测出的第一参考用户对第一参考物品进行操作的概率精度更高，提高了模型收敛速度。

在一些可能的实施方式中，根据第一训练样本和多个第二训练样本的相似度从所述多个第二训练样本中获取多个目标训练样本之前，方法还包括：将所述第一训练样本的多个所述用户特征域数据和多个所述物品特征域数据作为元素，对所述多个训练样本进行倒排索引，得到所述多个第二训练样本。

可选的，基于每个训练样本的用户特征域数据和物品特征域数据对多个训练样本进行倒排列表，其中，倒排列表包含有元素和样本之间的对应关系，如表2所示，倒排列表中的每一行中的第一列为元素，即样本下的一个域数据（用户特征域数据或者物品特征域数据），第二列为多个参考样本中包含有该域数据的参考样本。有了倒排列表之后，将第一训练样本中的每个用户特征域数据和每个物品特征域数据作为元素从多个训练样本中索引出多个第二训练样本，即根据倒排列表中的对应关系，可获取与每个用户特征域数据对应的训练样本，以及与每个物品特征域对应的训练样本；然后，将每个用户特征域数据对应的训练样本，以及与每个物品特征域对应的训练样本进行合并与去重，得到该多个第二训练样本。

可以看出，在本实施方式中，通过倒排索引对多个训练样本进行排序，得到倒排列表。由于使用了倒排列表，从而可以使用倒排列表快速的找到多个第二训练样本，不用和每个训练样本计算相似度，减轻了计算压力，可以快速从多个第二训练样本中获取出多个目标训练样本，提高模型训练速度。

第三方面，本申请实施例提供了一种推荐装置，包括：获取单元和处理单元；获取单元，用于获取待预测数据；处理单元，用于根据待预测数据和多个参考样本的相似度从多个参考样本中获取多个目标参考样本；每个参考样本和待预测数据均包括用户特征域数据和物品特征域数据，待预测数据的用户特征域数据用于指示目标用户特征，待预测数据的物品特征域数据用于指示目标物品特征，每个目标参考样本和待预测数据具有部分相同的用户特征域数据和/或物品特征域数据；根据多个目标参考样本与待预测数据获取待预测数据的目标特征信息；目标特征信息包括第一目标特征向量组和第二目标特征向量组，第一目标特征向量组为向量化后的待预测数据，第二目标特征向量组为对多个目标参考样本进行向量化后融合得到；以目标特征信息为输入通过深度神经网络DNN获取输出值；根据输出值确定是否向目标用户推荐目标物品。

在一些可能的实施方式中，多个目标参考样本还包括标签数据；第二目标特征向量组为

对多个目标参考样本进行向量化后融合得到，具体为：第二目标特征向量组为对多个目标参考样本的用户特征域数据，物品特征域数据以及标签数据进行向量化后融合得到。

在一些可能的实施方式中，目标特征信息还包括第三目标特征向量组，第三目标特征向量组为对第一向量组中的目标特征向量进行两两交互得到，第一向量组包括第一目标特征向量组和第二目标特征向量组。

在一些可能的实施方式中，在处理单元进行融合方面，处理单元，具体用于：对第一目标特征向量组中的多个第一目标特征向量进行拼接，得到待预测数据的第二特征向量；对每个目标参考样本的多个第一特征向量进行拼接，得到每个目标参考样本的第二特征向量，每个目标参考样本的多个第一特征向量为对目标参考样本进行向量化得到；获取每个目标参考样本的第二特征向量与待预测数据的第二特征向量之间的相似度；根据每个目标参考样本的第二特征向量与待预测数据的第二特征向量之间的相似度，确定每个目标参考样本的权重；根据每个目标参考样本的权重，对多个目标参考样本在同一个特征域下的第一特征向量进行融合，得到第二目标特征向量组。

在所述处理单元根据所述待预测数据和多个参考样本的相似度从所述多个参考样本中获取多个目标参考样本之前，所述处理单元，还用于：获取多个原始样本，其中，每个所述原始样本包括用户特征域数据和物品特征域数据；

将所述待预测数据的多个所述用户特征域数据和多个所述物品特征域数据作为元素，对所述多个原始样本进行倒排索引，得到所述多个参考样本。

第四方面，本申请实施例提供了一种推荐模型训练装置，推荐模型包括特征信息提取网络和深度神经网络 DNN，装置包括：获取单元和处理单元；获取单元，用于获取多个训练样本，其中，每个训练样本包括用户特征域数据和物品特征域数据；处理单元，用于根据第一训练样本和多个第二训练样本的相似度从多个第二训练样本中获取多个目标训练样本，其中，第一训练样本为多个训练样本中的一个，多个第二训练样本为多个训练样本除第一训练样本之外的部分或全部，第一训练样本的用户特征域数据用于指示第一参考用户特征，第一训练样本的物品特征域数据用于指示第一参考物品特征，第一训练样本和每个目标训练样本具有部分相同的用户特征域数据和/或物品特征域数据；将第一训练样本和多个目标训练样本输入到特征信息提取网络，得到第一训练样本的目标特征信息，其中，目标特征信息包括第四目标特征向量组和第五目标特征向量组，第四目标特征向量组为通过特征信息提取网络对第一训练样本进行向量化得到，第五目标特征向量组为通过特征信息提取网络对多个目标训练样本进行向量化后融合得到；将目标特征信息输入到深度神经网络 DNN，得到输出值，输出值用于表征第一参考用户对第一参考物品进行操作的概率；根据输出值以及第一训练样本的标签数据进行推荐模型的训练，获得目标推荐模型。

在一些可能的实施方式中，第五目标特征向量组为对多个目标训练样本进行向量化后融合得到，具体为：第五目标特征向量组为通过特征信息提取网络对多个目标训练样本的用户特征域数据、物品特征域数据以及标签数据进行向量化后融合得到。

在一些可能的实施方式中，目标特征信息还包括第六目标特征向量组，第六目标特征向量组是通过特征信息提取网络对第二向量组中的目标特征向量进行两两交互得到，第二向量组包括第四目标特征向量组和第五目标特征向量组。

在一些可能的实施方式中，在处理单元进行融合方面，处理单元，具体用于：对第四目标特征向量组中的多个第四目标特征向量进行拼接，得到第一训练样本的第二特征向量；对每个目标训练样本的多个第一特征向量进行拼接，得到每个目标训练样本的第二特征向量，

每个目标训练样本的多个第一特征向量为对目标训练样本进行向量化得到；获取每个目标训练样本的第二特征向量与第一训练样本的第二特征向量之间的相似度；根据每个目标训练样本的第二特征向量与第一训练样本的第二特征向量之间的相似度，确定每个目标训练样本的权重；根据每个目标训练样本的权重，对多个目标训练样本在同一个特征域下的第一特征向量进行融合，得到第五目标特征向量组。

在一些可能的实施方式中，在所述处理单元根据所述待预测数据和多个参考样本的相似度从所述多个参考样本中获取多个目标参考样本之前，所述处理单元，还用于：

将每个训练样本的每个用户特征域数据和每个物品特征域数据作为元素，对多个训练样本进行倒排索引，得到倒排列表；将第一训练样本的每个用户特征域数据和每个物品特征域数据作为查询词，从倒排列表中获取多个第二训练样本。

在一些可能的实施方式中，在所述处理单元根据第一训练样本和多个第二训练样本的相似度从所述多个第二训练样本中获取多个目标训练样本之前，所述处理单元，还用于：

将所述第一训练样本的多个所述用户特征域数据和多个所述物品特征域数据作为元素，对所述多个训练样本进行倒排索引，得到所述多个第二训练样本。

第五方面，本申请实施例提供了一种电子设备，包括：存储器，用于存储程序；处理器，用于执行存储器存储的程序；当存储器存储的程序被执行时，处理器用于实现上述第一方面或第二方面中的方法。

第六方面，本申请实施例提供了提供一种计算机可读介质，该计算机可读介质存储用于设备执行的程序代码，该程序代码包括用于实现上述第一方面或第二方面中的方法。

第七方面，本申请实施例提供了提供一种包含指令的计算机程序产品，当该计算机程序产品在计算机上运行时，使得计算机实现上述第一方面或第二方面中的方法。

第八方面，本申请实施例提供了提供一种芯片，该芯片包括处理器与数据接口，处理器通过数据接口读取存储器上存储的指令，实现上述第一方面或第二方面中的方法。

可选地，作为一种实现方式，芯片还可以包括存储器，存储器中存储有指令，处理器用于执行存储器上存储的指令，当指令被执行时，处理器用于实现上述第一方面或第二方面中的方法。

附图说明

图 1 为本申请实施例提供的一种人工智能主体框架示意图；

图 2 为本申请实施例提供的一种系统架构的示意图；

图 3 为本申请实施例提供的一种芯片硬件结构图；

图 4 为本申请实施例提供的一种推荐方法的流程示意图；

图 5 为本申请实施例提供的一种特征向量交互与拼接的示意图；

图 6 为本申请实施例提供的一种模型的结构图；

图 7 为本申请实施例提供的一种推荐模型训练方法的流程示意图；

图 8 为本申请实施例提供的一种用户行为预测流程比对图；

图 9 为本申请实施例提供的一种应用程序推荐的示意图；

图 10 为本申请实施例提供的一种商品推荐的示意图；

图 11 为本申请实施例提供的一种歌曲推荐的示意图；

图 12 为本申请实施例提供的一种推荐装置的结构图；

图 13 为本申请实施例提供的一种推荐模型训练装置的结构图；

图 14 为本申请实施例提供的一种电子设备的结构图。

具体实施方式

下面将结合本发明实施例中的附图，对本发明实施例中的技术方案进行描述，显然，所描述的实施例仅仅是本发明一部分实施例，而不是全部的实施例。基于本发明中的实施例，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，都属于本发明保护的范围。

参阅图 1，图 1 为本申请实施例提供的一种人工智能主体框架示意图。该主体框架描述了人工智能系统总体工作流程，适用于通用的人工智能领域需求。

下面从“智能信息链”（水平轴）和“IT 价值链”（垂直轴）两个维度对上述人工智能主题框架进行阐述。

“智能信息链”反映从数据的获取到处理的一列过程。举例来说，可以是智能信息感知、智能信息表示与形成、智能推理、智能决策、智能执行与输出的一般过程。在这个过程中，数据经历了“数据—信息—知识—智慧”的凝练过程。

“IT 价值链”从人智能的底层基础设施、信息（提供和处理技术实现）到系统的产业生态过程，反映人工智能为信息技术产业带来的价值。

（1）基础设施：

基础设施为人工智能系统提供计算能力支持，实现与外部世界的沟通，并通过基础平台实现支撑。通过传感器与外部沟通；计算能力由智能芯片提供，比如，智能芯片可以为中央处理器（central processing unit，CPU）、神经网络处理器（Neural-network Processing Unit ，NPU）、图形处理器（英语：graphics processing unit，缩写：GPU）、专用集成电路（Application SpecificIntegrated Circuit，ASIC）、现场可编程逻辑门阵列（Field Programmable Gate Array，FPGA）等硬件加速芯片；基础平台包括分布式计算框架及网络等相关的平台保障和支持，可以包括云存储和计算、互联互通网络等。举例来说，传感器和外部沟通获取数据，这些数据提供给基础平台提供的分布式计算系统中的智能芯片进行计算。

（2）数据：

基础设施的上一层的数据用于表示人工智能领域的数据来源。数据涉及到图形、图像、语音、文本，还涉及到传统设备的物联网数据，包括已有系统的业务数据以及力、位移、液位、温度、湿度等感知数据。

（3）数据处理：

数据处理通常包括数据训练，机器学习，深度学习，搜索，推理，决策等方式。

其中，机器学习和深度学习可以对数据进行符号化和形式化的智能信息建模、抽取、预处理、训练等。

推理是指在计算机或智能系统中，模拟人类的智能推理方式，依据推理控制策略，利用形式化的信息进行机器思维和求解问题的过程，典型的功能是搜索、匹配以及预测。

决策是指智能信息经过推理后进行决策的过程，通常提供分类、排序、预测等功能。

（4）通用能力：

对数据经过上面提到的数据处理后，进一步基于数据处理的结果可以形成一些通用的能力，比如可以是算法或者一个通用系统，例如，翻译，文本的分析，用户行为预测，计算机

视觉的处理，语音识别，图像的识别等等。

（5）智能产品及行业应用：

智能产品及行业应用指人工智能系统在各领域的产品和应用，是对人工智能整体解决方案的封装，将智能信息决策产品化、实现落地应用，其应用领域主要包括：智能制造、智能交通、智能家居、智能医疗、智能安防、自动驾驶，智能终端等。

参阅图 2，图 2 为本申请实施例提供的一种系统架构 200 的示意图。数据采集设备 260 用于采集包括用户特征域数据、物品特征域数据以及标签数据的多域离散数据，即训练样本，并将训练样本存入数据库 230，训练设备 220 基于数据库 230 中维护的训练样本生成模型/规则 201。下面将更详细地描述训练设备 220 如何基于训练样本得到模型/规则 201，模型/规则 201 能够对待预测数据进行处理，得到输出值，即目标用户对目标物品进行操作的概率，以便根据输出值确定是否向目标用户推荐目标物品。

训练设备 220 得到的模型/规则可以应用不同的系统或设备中。在附图 2 中，执行设备 210 配置有 I/O 接口 212，与外部设备进行数据交互，“用户”可以通过客户设备 240 向 I/O 接口 212 输入数据，比如，可以通过客户设备 240 向 I/O 接口 212 输入待预测数据，其中，待预测数据中包括用户特征域数据和物品特征域数据，“用户”向执行设备 210 输入待预测数据的目的是获取输出值，以得到目标用户对目标物品进行操作的概率。

执行设备 210 可以调用数据存储系统 250 中存储的数据、代码等，也可以将数据、指令等存入数据存储系统 250 中。其中，数据存储系统 250 中存储有大量的参考样本，参考样本可以是数据库 230 中维护的训练样本，即数据库 230 可以将数据迁移到数据存储系统 250；

关联功能模块 213 对待预测数据进行分析，从数据存储系统 250 中维护的参考样本中查询出多个目标参考样本；

计算模块 211 使用模型/规则 201 对关联功能模块 213 查询出的多个目标参考样本以及待预测数据进行处理。具体的，计算模块 211 调用模型/规则 201 对多个目标参考样本进行向量化后融合，以及对待预测数据进行向量化处理，得到待预测数据的目标特征信息，并根据目标特征信息得到输出值；

最后，计算模块 211 通过 I/O 接口 212 将输出值返回给客户设备 240，从而使客户设备 240 获取到目标用户对目标物品进行操作的概率。

更深层地，训练设备 220 可以针对不同的目的，基于不同的数据生成相应的模型/规则 201，以给用户提供更佳的结果。

在图 2 中所示的情况下，用户可以手动指定输入执行设备 210 中的数据，例如，在 I/O 接口 212 提供的界面中操作。另一种情况下，客户设备 240 可以自动地向 I/O 接口 212 输入数据并获得结果，如果客户设备 240 自动输入数据需要获得用户的授权，用户可以在客户设备 240 中设置相应权限。用户可以在客户设备 240 查看执行设备 210 输出的结果，具体的呈现形式可以是显示、声音、动作等具体方式。客户设备 240 也可以作为数据采集端将采集到数据存入数据库 230。

值得注意的是，图 2 仅是本发明实施例提供的一种系统架构的示意图，图 2 中所示设备、器件、模块等之间的位置关系不构成任何限制，例如，在图 2 中，数据存储系统 250 相对执行设备 210 是外部存储器，在其它情况下，也可以将数据存储系统 250 置于执行设备 210 中。

参阅图 3，图 3 是本申请实施例提供的一种芯片硬件结构图。神经网络处理器

（Neural-network Processing Unit，NPU）30 作为协处理器挂载到主中央处理器（Central Processing Unit，CPU）上，由主 CPU 分配任务。NPU 的核心部分为运算电路 303，控制器 304 控制运算电路 303 提取存储器（权重存储器 302 或输入存储器 301）中的数据并进行运算。

在一些实现中，运算电路 303 内部包括多个处理单元（Process Engine, PE）。

在一些实现中，运算电路 303 是二维脉动阵列。运算电路 303 还可以是一维脉动阵列或者能够执行例如乘法和加法这样的数学运算的其它电子线路。

在一些实现中，运算电路 303 是通用的矩阵处理器。

举例来说，假设有输入矩阵 A，权重矩阵 B，输出矩阵 C。运算电路 303 从权重存储器 302 中取权重矩阵 B，并缓存在运算电路 303 中每一个 PE 上。运算电路 303 从输入存储器 301 中取输入矩阵 A 与权重矩阵 B 进行矩阵运算，得到的矩阵的部分结果或最终结果，保存在累加器（accumulator）308 中。

向量计算单元 307 可以对运算电路 303 的输出做进一步处理，如向量乘，向量加，指数运算，对数运算，大小比较等等。例如，向量计算单元 307 可以用于神经网络中非卷积/非（Fully Connected Layers，FC）层的网络计算，如池化（Pooling），批归一化（Batch Normalization），局部响应归一化（Local Response Normalization）等。

在一些实现中，向量计算单元 307 将经处理的向量存储到统一缓存器 306。例如，向量计算单元 307 可以将非线性函数应用到运算电路 303 的输出，例如累加值的向量，用以生成激活值。在一些实现中，向量计算单元 307 生成归一化的值、合并值，或二者均有。在一些实现中，处理过的向量能够用作到运算电路 303 的激活输入，例如用于在神经网络中的后续层中的使用。

示例性的，对于本申请来说，运算电路 303 从输入存储器 301 中获取待预测数据，从统一存储器 306 中获取目标参考样本；然后，运算电路 303 根据待预测数据以及对目标参考样本获取待预测数据的目标特征信息，并根据目标特征信息得到输出值，即目标用户对目标物品进行操作的概率。

统一存储器 306 用于存放输入数据（比如，待预测数据）以及输出数据（比如，输出值）。

存储单元访问控制器（Direct Memory Access Controller，DMAC）305 将外部存储器中的输入数据搬运到输入存储器 301 和/或统一存储器 306、将外部存储器中的权重数据存入权重存储器 302，以及将统一存储器 306 中的数据存入外部存储器。

总线接口单元（Bus Interface Unit，BIU）310，用于通过总线实现主 CPU、DMAC 和取指存储器 309 之间进行交互。

取指存储器（instruction fetch buffer）309，用于存储控制器 304 使用的指令；

控制器 304 用于调用取指存储器 309 中缓存的指令，实现控制运算电路 303 的工作过程。

一般地，统一存储器 306，输入存储器 301，权重存储器 302 以及取指存储器 309 均为片上（On-Chip）存储器，外部存储器为 NPU 外部的存储器，外部存储器可以为双倍数据率同步动态随机存储器（Double Data Rate Synchronous Dynamic Random Access Memory，DDR SDRAM）、高带宽存储器（High Bandwidth Memory，HBM）或其他可读可写的存储器。

为了便于理解本申请，下面介绍与本申请相关的概念。

表格式数据（Tabular Data），又称为多域离散数据(Multi-Field Categorical Data)，如表 1 所示，表格式数据中的每行为一个数据点（又称作一个样本），每列代表一个特征(又称作一个域，也可称作特征域)，则每个样本包含有多个特征域；并且，样本在每个特征域下的取值

称为特征域数据，也可称为域值。比如，表 1 中的 LA、NYC、LA、London 分别为样本 1、样本 2、样本 3 以及样本 4 在城市这个特征域下的域值。

表 1：

样本	用户 ID	城市	身份	物品 ID	种类	商标
样本 1	U1	LA	Student	L1	T-shirt	B1
样本 2	U2	NYC	Student	L1	T-shirt	B1
样本 3	U1	LA	Student	L2	cell phone	B2
样本 4	U3	London	manager	L3	cell phone	B3

当表格式数据用于预测用户行为时，则每个样本的特征域包括用户特征域和物品特征域，且在用户特征域下的域值称为用户特征域数据，在物品特征域下的域值称为物品特征域数据。一般来说，用户特征域数据包括用户的属性信息和用户的行为序列（可选的），其中，用户的属性信息包括用户的标识（ID）、居住地、身份、性别、年龄，等其他基本信息，物品特征域数据包括物品的 ID、种类、商标、尺寸、颜色，等其他基本信息；用户的行为序列包括用户的历史行为，比如，用户过去点击、浏览、购买过的物品，等等。

参阅图 4，图 4 为本申请实施例提供的一种推荐方法的流程示意图。该方法包括以下步骤内容：

401：获取待预测数据。

待预测数据为多域离散数据，且待预测数据包含有用户特征域数据和物品特征域数据。

可选的，待预测数据的用户特征域数据用于指示目标用户特征。示例性的，该用户特征域数据包括目标用户的属性信息，比如，目标用户的 ID、年龄、性别、居住地、户籍地，等其他基本信息；

可选的，待预测数据的物品特征域数据用于指示目标物品特征。其中，目标物品可以为商品、应用程序、歌曲、网页，等与用户相关的物品。对于不同的目标物品，目标物品特征域数据可以有不同的表现。比如，目标物品为应用程序，则目标物品特征域数据包括应用程序的类型、安装大小、访问热度、安装次数，等等；再如，目标物品为歌曲，则目标物品特征域数据包括歌曲的风格、节奏、时长、播放次数、播放热度，等等；又如，目标物品为商品，目标物品特征域数据包括商品的颜色、尺寸、价格、商标、生产厂家、评价，等等。

可选的，待预测数据的用户特征域数据中还可以包含目标用户的行为序列，比如，目标用户的行为序列包括目标用户过去点击、浏览、购买过的物品，等等。

402：根据待预测数据和多个参考样本的相似度从多个参考样本中获取多个目标参考样本。

示例性的，获取待预测数据和多个参考样本中每个参考样本之间的相似度，根据待预测数据和每个参考样本之间的相似度，从多个参考样本中获取出多个目标参考样本。

其中，每个参考样本也为多域离散数据，每个参考样本也包括用户特征域数据和物品特征域数据。其中，每个参考样本的用户特征域数据用于指示参考用户特征，物品特征域数据用于指示参考物品特征。与待预测数据类似的，每个参考样本的用户特征域数据包含参考用户的属性信息，物品特征域数据包含参考物品的属性信息，比如，颜色、形状、价格，等其他信息，不再叙述。

其中，每个目标参考样本和待预测数据具有部分相同的用户特征域数据和/或物品特征域数据。应说明，为了确保目标参考样本对待预测数据真正的具有参考作用，需要使每个目标参考样本和待预测数据具有部分相同的用户特征域数据和物品特征域数据。举例来说，如果

目标参考样本和待预测数据只是具有部分相同的用户特征域数据，比如，都是男性，这种目标参考样本对待预测数据的行为预测并不具有参考价值；又或者，目标参考样本和待预测数据只是具有部分相同的物品特征域数据，比如，购买的物品都是黑色，这种目标参考样本对待预测数据的行为预测也是不具有参考价值。因此，在实际应用中，所获得的目标参考样本和待预测数据相比，需要同时具有部分相同的用户特征域数据和物品特征域数据。

一般来说，每个参考样本包含的用户特征域数据和待预测数据包含的用户特征域数据不完全相同，而每个参考样本包含的物品特征域数据与待预测数据包含的物品域特征数据可以完全相同。

应说明，多个参考样本可以按照表 1 示出的方式预先存储好，形成表格式数据，也可以自由存储，只要这些参考样本的多个特征域和待预测数据的多个特征域相同即在本申请的保护范围内，不限定对多个参考样本的存储方式。

示例性的，该多个参考样本可以为样本库中的多个原始样本，也可以是从多个原始样本中选出的样本，其中，每个原始样本也为多域离散数据，与待预测数据类似的，每个原始样本也包括用户特征域数据和物品特征域数据，不再叙述。

可选的，若多个参考样本是从多个原始样本中选出的，则为了从多个原始样本中快速获取到多个参考样本，可对多个原始样本进行倒排索引，得到倒排列表，基于倒排列表获取多个参考样本。

示例性的，将每个原始样本的每个用户特征域数据和每个物品特征域数据均作为元素（item），以及将每个原始样本作为文档（document），对多个原始样本进行倒排索引，得到倒排列表。由于本申请只需要从多个原始样本中获取多个参考样本，不关注元素在文档中出现的次数和其他信息，因此本申请中的倒排列表可以只包含有元素和文档的对应关系。

因此，通过倒排索引可以将表 1 示出的多个原始样本转化为表 2 所示的倒排列表：

表 2：

元素	文档
U1	样本 1、样本 3
U2	样本 2
U3	样本 4
LA	样本 1、样本 3
NYC	样本 2
London	样本 4
Student	样本 1、样本 2、样本 3
L1	样本 1、样本 2
L2	样本 3
L3	样本 4
T-shirt	样本 1、样本 2
cell phone	样本 3、样本 4
B1	样本 1、样本 2
B2	样本 3
B3	样本 4

然后，将待预测数据的每个用户特征域数据和每个物品特征域数据均作为元素，从倒排

列表中索引出与待预测数据对应的多个参考样本。即从倒排列表中索引出与待预测数据的每个用户特征域数据对应的参考样本，以及与每个物品特征域对应的参考样本；然后将索引出的所有参考样本进行合并与去重，得到多个参考样本。因此，每个参考样本与待预测数据相比，至少在一个相同特征域下的域数据相同，比如，在同一个用户特征域下的域数据相同，比如，居住在同一个城市。

举例来说，待预测数据为[U4、LA、Student、L2、cell phone、B3]，将U4、LA、Student、L2、cell phone、B3均作为查询词，从倒排列表中索引出与LA对应的参考样本[样本1、样本3]，与Student对应的参考样本[样本1、样本2、样本3]，与L2对应的参考样本为[样本3]，与cell phone对应的参考样本[样本3、样本4]，与B3对应的参考样本[样本4]。然后，对所有的参考样本进行合并与去重之后，得到多个参考样本[样本1、样本2、样本3、样本4]。

可以看出，通过倒排的方式先对原始样本进行存储，从而可以先从多个原始样本中索引出一部分原始样本作为多个参考样本，这样后续只需要计算待预测数据和多个参考样本的相似度，而不用计算和多个原始样本的相似度，减轻计算压力，从而可以快速获取出多个目标参考样本。

进一步的，得到多个参考样本之后，可获取待预测数据与每个参考样本之间的相似度。可选的，通过BM25算法获取待预测数据与每个参考样本之间的相似度，不再叙述。

示例性的，将相似度大于阈值的参考样本作为目标参考样本，得到多个目标参考样本，或者，按照相似度从高到低的顺序从多个参考样本中选取预设数量的参考样本，作为多个目标参考样本。

403：根据多个目标参考样本与待预测数据获取待预测数据的目标特征信息。

可选的，目标特征信息包括第一目标特征向量组和第二目标特征向量组，其中，第一目标特征向量组为向量化后的待预测数据，第二目标特征向量组为对多个目标参考样本进行向量化后融合得到。

示例性的，对待预测数据进行向量化，得到第一目标特征向量组，其中，第一目标特征向量组包括多个第一目标特征向量。

可选的，对待预测数据的每个用户特征域数据和每个物品特征域数据均进行编码，得到待预测数据的特征向量。对待预测数据的每个用户特征域数据和每个物品特征域数据均进行编码可以理解为将待预测样本待预测数据的每个用户特征域数据和每个物品特征域数据均进行数字化处理，得到待预测数据的特征向量；然后，对待预测数据的特征向量进行映射处理，得到多个目标第一特征向量，其中，每个第一目标特征向量用于表示待预测数据的一个特征域数据，即将待预测数据中每个特征域数据的编码结果进行映射，得到与该特征域数据对应的第一目标特征向量。

应说明，若待预测数据中包含有目标用户的行为序列，则将目标用户的行为序列进行编码，并将编码结果进行映射，得到映射结果；然后，将用户行为序列对应的映射结果进行融合，得到一个与目标用户的行为序列对应的第一目标特征向量，该第一目标特征向量用于表示目标用户的行为序列。

可选的，获取多个第一目标特征向量可以通过目标推荐模型实现，后面详细描述目标推荐模型的训练过程，在此不做过多描述。

具体的，目标推荐模型包括特征信息提取网络和深度神经网络（Deep Neural Networks，DNN），其中，该DNN可以为多层感知器（Multi-Layer Perceptron，MLP），并且本申请中以DNN为MLP为例进行说明，不再赘述。其中，该特征信息提取网络包括编码层和映射层

（embedding layer）。将该待预测数据输入到编码层对待预测数据的每个用户特征域数据和每个物品特征域数据均进行编码，得到待预测数据的特征向量（$c_1, c_2, c_3, \ldots, c_n$），其中，$c_1, c_2, c_3, \ldots, c_n$分别表示待预测数据的第 1 个，第 2 个，第 3 个，…，第 n 个特征域数据的编码结果；然后，将特征向量（$c_1, c_2, c_3, \ldots, c_n$）输入到映射层进行映射，得到多个第一目标特征向量（$e_1, e_2, e_3, \ldots, e_n$），即将待预测数据的第 1 个，第 2 个，第 3 个，…，第 n 个特征域数据的编码结果分别进行映射，得到多个第一目标特征向量（$e_1, e_2, e_3, \ldots, e_n$）。

针对参考样本来说，可以携带标签数据也可以不携带标签数据。下面分别叙述参考样本携带有标签和不携带标签数据获取第二目标特征向量组的过程。

针对不携带标签数据的情况：

示例性的，对每个目标参考样本进行向量化，得到每个目标参考样本的多个第一特征向量。可选的，对每个目标参考样本的每个用户特征域数据和每个物品特征域数据均进行编码，得到每个目标参考样本的特征向量；对每个目标参考样本的特征向量进行映射处理，得到每个目标参考样本的多个第一特征向量，其中，每个第一特征向量用于表示该目标参考样本的一个特征域数据。

然后，对多个目标参考样本的第一特征向量进行融合，得到第二目标特征向量组，其中，第二目标特征向量组包括多个第二目标特征向量。

示例性的，根据待预测数据的多个第一目标特征向量以及每个目标参考样本的多个第一特征向量，确定每个目标参考样本对应的权重；根据多个目标参考样本的多个权重，对多个目标参考样本在同一个特征域下的第一特征向量进行融合，得到多个第二目标特征向量。

具体的，将待预测数据的多个第一目标特征向量进行拼接，得到待预测数据的第二特征向量；将每个目标参考样本的多个第一特征向量进行拼接，得到每个目标参考样本的第二特征向量；获取待预测数据的第二特征向量与每个目标参考样本的第二特征向量之间的相似度，得到与多个目标参考样本对应的多个相似度，其中，相似度可以为欧氏距离、余弦相似度，等等；然后，对多个目标参考样本的多个相似度进行归一化处理，并将每个目标参考样本对应的归一化结果作为每个目标参考样本的权重。因此，目标参考样本中的第 i 个目标参考样本的权重可以通过公式（1）表示：

$$a_i = \frac{similarity\ (q,\ r_i)}{\sum_{i=1}^{k} similarity\ (q,\ r_i)} \quad \text{公式（1）；}$$

其中，a_i为第 i 个目标参考样本的权重，q 为待预测数据的第二特征向量，r_i为第 i 个目标参考样本的第二特征向量，$similarity\ (q,\ e_i)$为第 i 个目标参考样本的第二特征向量与待预测数据的第二特征向量之间的相似度，k 为多个目标参考样本的数量。

可选的，根据每个目标参考样本的权重，对多个目标参考样本在任意一个相同特征域下的第一特征向量（即每个目标参考样本在该特征域下对应一个第一特征向量）进行融合，即加权处理，得到该特征域下的第二目标特征向量；然后，对多个目标参考样本中在每个相同域下的第一特征向量分别进行融合，得到多个第二目标特征向量，所以多个第二目标特征向量的数量与目标参考样本的特征域的数量相同。示例性的，多个第二目标特征向量中的第 j 个第二目标特征向量可以通过公式（2）表示：

$$R_j = \sum_{i=1}^{k} a_i * e_{ij} \quad \text{公式（2）；}$$

其中，R_j为第 j 个第二目标特征向量，e_{ij}为第 i 个目标参考样本的第 j 个第一特征向量，j 的取值为 1 到 n 的整数，n 为每个目标参考样本的多个第一特征向量的数量，即每个目标参考样本的特征域的数量，也是多个第二目标特征向量的数量。

可选的，由于已经将每个目标参考样本的多个第一特征向量进行拼接，得到了每个目标

参考样本的第二特征向量。因此，在得到每个目标参考样本的第二特征向量之后，可以使用每个目标参考样本的权重对多个目标参考样本的多个第二特征向量直接进行融合，即加权处理，得到一个融合后的特征向量；然后，按照对每个目标参考样本的多个第一特征向量进行拼接的反顺序，对融合后的特征向量进行拆分，得到多个第二目标特征向量。

应理解，在对待预测数据的多个第一特征向量以及每个目标参考样本的第一特征向量进行拼接时，并不限定对多个第一特征向量进行拼接的顺序，但是需要保证对待预测数据的多个第一目标特征向量的拼接顺序和对每个目标参考样本的多个第一特征向量的拼接顺序是一致的。

可选的，获取多个第二目标特征向量也可以通过上述的目标推荐模型实现。

示例性的，将每个目标参考样本的用户特征域数据和物品特征域数据输入到编码层对每个目标参考样本进行编码，得到每个目标参考样本的特征向量，比如，第 i 个目标参考样本的特征向量为（$r_{i1}, r_{i2}, r_{i3}, \ldots, r_{in}$），i 的取值为 1 到 k，k 为多个目标参考样本的数量；然后，将每个目标参考样本的特征向量输入到嵌入层对每个目标参考向量的特征向量进行映射处理，得到每个目标参考样本的多个第一特征向量，比如，第 i 个目标参考样本的多个第一特征向量为（$e_{i1}, e_{i2}, e_{i3}, \ldots, e_{in}$）；可选的，特征信息提取网络还包括注意力层，将每个目标参考样本的多个第一特征向量以及待预测数据的多个第一目标特征向量（$e_1, e_2, e_3, \ldots, e_n$）输入到注意力层，将（$e_1, e_2, e_3, \ldots, e_n$）进行拼接得到待预测数据的第二特征向量，即$\begin{pmatrix} e_1 \\ e_2 \\ \cdots \\ e_n \end{pmatrix}$；将每个目标参考样本的第一特征向量进行拼接，得到每个目标参考样本的第二特征向量，比如，第 i 个目标参考样本的第二特征向量为$\begin{pmatrix} e_{i1} \\ e_{i2} \\ \cdots \\ e_{in} \end{pmatrix}$；然后，基于待预测数据的第二特征向量以及每个目标参考样本的第二特征向量，确定每个目标参考样本的权重；最后，基于每个目标参考样本的权重对多个目标参考样本的第一特征向量进行融合，得到多个第二目标特征向量，即（$a_{1*}e_{11} + a_{2*}e_{21} + \cdots + a_{k*}e_{k1}, a_{1*}e_{12} + a_{2*}e_{22} + \cdots + a_{k*}e_{k2}, \ldots, a_{1*}e_{1n} + a_{2*}e_{2n} + \cdots + a_{k*}e_{kn}$），其中，该多个第二目标特征向量还可以简化表示为：（$e_{n+1}, e_{n+2}, e_{n+3}, \ldots, e_{2n}$）。

针对携带标签数据的情况：

示例性的，每个参考样本还携带有标签数据，该标签数据用于表征参考样本中的参考用户对参考物品的实际操作情况，比如，参考物品为应用程序时，该标签用于表征参考用户是否点击了该应用程序。因此，在对每个目标参考样本进行向量化，得到每个目标参考样本的多个第一特征向量过程中，除了将每个目标参考样本的每个用户特征域数据和每个物品特征域数据进行向量化之外，还会同步将每个目标参考样本的标签数据进行向量化，得到每个目标参考样本的多个第一特征向量，因此与上述不携带标签数据的情况相比，此时向量化得到的每个目标参考样本的多个第一目标特征向量中还包含用于指示标签数据的第一目标特征向

量。具体的，将每个目标参考样本的每个用户特征域数据、每个物品特征域数据以及标签数据进行编码，得到每个目标参考样本的特征向量。比如，第 i 个目标参考样本的特征向量为$(r_{i1}, r_{i2}, r_{i3}, \ldots, r_{in}, r_{i(n+1)})$，其中，$r_{i(n+1)}$为第 i 目标参考样本的标签数据的编码结果；然后，对每个目标参考样本的特征向量进行映射，得到每个目标参考样本的多个第一特征向量。比如，第 i 个目标参考样本的多个特征向量为$(e_{i1}, e_{i2}, e_{i3}, \ldots, e_{in}, e_{i(n+1)})$，其中，$e_{i(n+1)}$用于指示第 i 个目标参考样本的标签数据。

进一步的，与上述融合类似，根据上述计算出的每个目标参考样本的权重，对多个目标参考样本在同一个特征域（包括用户特征域、物品特征域以及标签域）下的第一特征向量进行融合，得到第二目标特征向量组，即与上述不携带标签数据的情况相比，此时融合得到的第二特征向量组中的多个第二特征向量中还包含用于指示融合后的标签数据的第二目标特征向量。举例来说，第二目标特征向量组为$(e_{n+1}, e_{n+2}, e_{n+3}, \ldots, e_{2n}, e_{2n+1})$，其中，$e_{2n+1}$用于指示多个目标参考样本的融合后的标签数据。

可选的，针对上述不携带标签数据或携带标签数据的情况，在获取到第一目标特征向量组和第二目标特征向量组之后，可以将第一目标特征向量组和第二目标特征向量组进行拼接，得到目标特征信息，则目标特征信息为$(e_1, e_2, e_3 \ldots, e_n, e_{n+1}, e_{n+2}, e_{n+3}, \ldots, e_{2n})$或者$(e_1, e_2, e_3 \ldots, e_n, e_{n+1}, e_{n+2}, e_{n+3}, \ldots, e_{2n}, e_{2n+1})$；可选的，也可以不对第一目标特征向量组和第二目标特征向量组进行拼接。比如，可以将第一目标特征向量组和第二目标特征向量组均作为输入数据进行后面的输出值的预测，得到输出值。

在本申请的一个实施方式中，在得到第一目标特征向量组和第二目标特征向量组之后，除了第一目标特征向量组和第二目标特征向量组进行拼接之外，还可以将目标特征向量进行交互，以使获得高阶的特征信息。示例性的，如图 5 所示，可以先将多个第一目标特征向量和多个第二目标特征向量进行拼接（concat），得到第一向量组；然后，对第一向量组中的目标特征向量进行两两交互，得到第三目标特征向量组；然后，将第一向量组与第三目标特征向量组进行拼接，得到目标特征信息。同样，也可以不对第一向量组与第三目标特征向量组进行拼接，将两者均作为输入数据即可，不再叙述。

示例性的，多个第三目标特征向量可以通过公式（3）表示：

$e_{ij} = \text{inter}(e_i, e_j)$ 公式（3）；

其中，i 的取值为 1 到 2n，j 的取值为 2 到 2n，且 j 的取值大于 i，2n 为第一向量组中的目标特征向量的数量，inter 为向量之间的交互操作。

其中，上述向量两两交互主要是将两个向量融合为一个向量，且融合后的一个向量所表示的特征信息为两个向量表示的特征信息所融合后的特征信息。可选的，向量两两交互可以通过向量点乘实现、核积实现以及网络层实现。本申请不限定两个向量的交互方式，只要能融合后得到一个向量能够表示两个向量所表示的特征信息即可。

应理解，上述只给出了将特征向量进行两两交互的情况，在实际应用中，还可以将三个特征向量或者更多数量的特征向量进行交互。另外，上述在进行向量交互的过程中，对第一向量组中的所有目标特征向量进行了两两交互，然而在实际应用中，也可以从第一向量组中选取部分目标特征向量进行交互，比如，可以只将目标特征信息中的部分多个第一目标特征向量和部分多个目标第二特征向量进行交互，得到多个第三目标特征向量。因此，本申请并不限定进行交互的向量的来源，以及进行交互的向量的数量。

404：以目标特征信息为输入通过深度神经网络 DNN 获取输出值。

示例性的，将目标特征信息作为输入数据输入到深度神经网络 DNN，得到输出值。

示例性的，一般来说输出值是一个概率值，该概率值表示目标用户对目标物品进行操作的概率。应说明，对于不同的目标物品，目标用户对目标物品进行操作的概率有不同的理解。比如，当目标物品为应用程序时，目标用户对目标物品进行操作的概率可以理解为目标用户对应用程序的点击概率；再如，当目标物品为歌曲时，目标用户对目标物品进行操作的概率可以理解为目标用户喜欢该歌曲的概率；又如，当目标物品为商品时，目标用户对目标物品进行操作的概率可以理解为目标用户购买该商品的概率。

实际应用中，获取了概率值之后，可以对概率值进行后处理，得到输出值，比如，当概率值大于概率阈值时，以 1 作为输出值，当概率值小于或等于阈值时，则以 0 作为输出值，其中，0 表示目标用户不会操作目标物品，1 表示目标用户会操作目标物品。

405：根据输出值确定是否向目标用户推荐目标物品。

可选的，当输出值是以 0 或 1 的二值化数据表示时，则当输出值为 1 时，确定向目标用户推荐目标物品，当输出值为 0 时，确定不向目标用户推荐目标物品。可选的，当输出值是以概率的形式表示时，则当概率大于概率阈值时，确定向目标用户推荐目标物品，当概率小于或等于概率阈值时，确定不向目标用户推荐目标物品。

应说明，当本申请的推荐方法应用到多物品的推荐场景时，则需要计算目标用户对每个候选物品的操作概率；然后，对多个候选物品的操作概率进行排序，将排名靠前的候选物品推荐给目标用户。例如，进行歌曲推荐时，需要计算目标用户对各个候选歌曲的喜欢概率，然后将喜欢概率靠前的歌曲推荐给目标用户。

可以看出，在本申请实施方式中，获取的目标特征信息中除了包含待预测数据本身的特征信息之外，还包含有多个目标参考样本向量化后融合得到的特征信息。由于目标参考样本和待预测数据具有部分相同的用户特征域数据和/或物品特征域数据，因此目标参考样本中的用户行为可以为目标用户的行为的预测提供参考和经验，从而在使用这样的目标特征信息进行输出值的预测时，可以使预测出的输出值比较精确，基于这个输出值进行物品推荐，提高了推荐的精度。

下面结合具体的模型结构，以及以参考样本携带有标签数据，且对目标特征向量进行交互的方式介绍获取输出值的过程。

如图 6 所示，模型包括特征信息提取网络以及 MLP，其中，特征信息提取网络包括编码层、嵌入层、注意力层和交互层。其中，交互层是可选的，当需要对目标特征向量进行交互时，需要设计交互层；若不对目标特征向量进行交互，则可以不设计交互层。

首先，将待预测数据输入到搜索引擎中，并从多个参考样本中获取 k 个目标参考样本，即 S_1、S_2、…、S_k。然后，将待预测数据的用户特征域数据和物品特征域数据输入到编码层进行编码，得到待预测数据的特征向量，即 $(c_1, c_2, c_3, \ldots, c_n)$；以及将每个目标参考样本的用户特征域数据、物品特征域数据以及标签数据输入到编码层进行编码，得到多个目标参考样本对应的多个特征向量，即 $(r_{11}, r_{12}, r_{13}, \ldots, r_{1n}, r_{1n+1})$，$(r_{21}, r_{22}, r_{23}, \ldots, r_{2n}, r_{2n+1})$，…，$(r_{k1}, r_{k2}, r_{k3}, \ldots, r_{kn}, r_{kn+1})$。然后，将 $(c_1, c_2, c_3, \ldots, c_n)$ 以及 $(r_{11}, r_{12}, r_{13}, \ldots, r_{1n}, r_{1n+1})$，$(r_{21}, r_{22}, r_{23}, \ldots, r_{2n}, r_{2n+1})$，…，$(r_{k1}, r_{k2}, r_{k3}, \ldots, r_{kn}, r_{kn+1})$ 分别输入到嵌入层，对 $(c_1, c_2, c_3, \ldots, c_n)$ 以及 $(r_{11}, r_{12}, r_{13}, \ldots, r_{1n}, r_{1n+1})$，$(r_{21}, r_{22}, r_{23}, \ldots, r_{2n}, r_{2n+1})$，…，$(r_{k1}, r_{k2}, r_{k3}, \ldots, r_{kn}, r_{kn+1})$ 进行映射处理，分别得到待预测数据的多个第一目标特征向量

$(e_1, e_2, e_3 \ldots, e_n)$，以及每个目标参考样本的多个第一特征向量，即$(e_{11}, e_{12}, e_{13} \ldots, e_{1n}, e_{1n+1})$，$(e_{21}, e_{22}, e_{23} \ldots, e_{2n}, e_{2n+1})$，…，$(e_{k1}, e_{k2}, e_{k3}, \ldots, e_{kn}, e_{kn+1})$；然后，将$(e_1, e_2, e_3, \ldots, e_n)$、$(e_{11}, e_{12}, e_{13} \ldots, e_{1n}, e_{1n+1})$，$(e_{21}, e_{22}, e_{23} \ldots, e_{2n}, e_{2n+1})$，…，$(e_{k1}, e_{k2}, e_{k3}, \ldots, e_{kn}, e_{kn+1})$均输入到注意力层，对$(e_{11}, e_{12}, e_{13} \ldots, e_{1n}, e_{1n+1})$，$(e_{21}, e_{22}, e_{23} \ldots, e_{2n}, e_{2n+1})$，…，$(e_{k1}, e_{k2}, e_{k3}, \ldots, e_{kn}, e_{kn+1})$进行融合，得到多个目标第二特征向量$(e_{n+1}, e_{n+2}, e_{n+3}, \ldots, e_{2n}, e_{2n+1})$；然后，分别将多个第一目标特征向量$(e_1, e_2, e_3, \ldots, e_n)$、多个第二目标特征向量$(e_{n+1}, e_{n+2}, e_{n+3}, \ldots, e_{2n}, e_{2n+1})$进行拼接（Concat），得到第一向量组，即$(e_1, e_2, e_3 \ldots, e_n, e_{n+1}, e_{n+2}, e_{n+3}, \ldots, e_{2n}, e_{2n+1})$；然后，对第一向量组中的目标特征向量进行两两交互，得到第三目标特征向量，即上述的inter(e_i, e_j)；将第一向量组中的目标特征向量与第三目标特征向量进行拼接，得到目标特征信息，即$(e_1, e_2, e_3 \ldots, e_n, e_{n+1}, e_{n+2}, e_{n+3} \ldots, e_{2n}, e_{2n+1}, \text{inter}(e_i, e_j))$；

最后，将目标特征信息输入到多层感知器MLP，得到输出值。

参阅图7，图7为本申请实施例提供的一种推荐模型训练方法的流程示意图。该推荐模型包括特征信息提取网络和多层感知器MLP。该方法包括以下步骤内容：

701：获取多个训练样本。

其中，每个训练样本为多域离散数据，与上述参考样本类似，每个训练样本包括用户特征域数据和用户特征域数据。应理解，每个训练样本还携带有标签数据，每个训练样本的标签数据用于表示每个训练样本中的用户对该训练样本中的物品的实际操作情况。比如，物品为应用程序时，实际操作情况为用户是否点击了该应用程序。

应说明，上述的多个参考样本可以是该多个训练样本，也可以是该多个训练样本中的部分训练样本，比如，从多个训练样本中筛选出部分数据完整度较高的训练样本作为参考样本。

702：根据第一训练样本和多个第二训练样本的相似度从多个第二训练样本中获取多个目标训练样本。

其中，第一训练样本为多个训练样本中的任意一个，第一训练样本的用户特征域数据用于指示第一参考用户特征，第一训练样本的物品特征域数据用于指示第一参考物品特征。其中，多个第二训练样本为多个训练样本中除第一训练样本之外的部分或全部。

第一训练样本和每个目标训练样本具有部分相同的用户特征域数据和/或物品特征域数据。同样的，在实际应用中，为了确保每个目标训练样本和第一训练样本的相似性，一般来说，每个目标参考样本和第一训练样本具有部分相同的用户特征域数据和物品特征域数据。

示例性的，根据第一训练样本与每个第二训练样本之间的相似度从多个第二训练样本中获取多个目标训练样本。比如，可以将相似度大于阈值的第二训练样本作为目标训练样本，得到多个目标训练样本，或者，按照相似度从高到低的顺序从多个第二训练样本中选取预设数量的第二训练样本，作为多个目标训练样本。

应理解，可以直接将多个训练样本中除第一训练样本之外的全部训练样本作为多个第二训练样本，然后获取第一训练样本与多个第二训练样本之间的相似度；也可以按照上述倒排索引的方式，从除第一训练样本之外的训练样本中选出部分作为多个第二训练样本。

示例性的，按照上述倒排索引的方式，将每个训练样本的每个用户特征域数据和每个物品特征域数据均作为元素，将每个训练样本作为文档，对多个训练样本进行倒排索引，得到倒排列表；然后，将第一训练样本的每个用户特征域数据和每个物品特征域数据均作为查询词，从倒排索引中获取出多个第二训练样本，因此，这些第二训练样本与第一训练样本相比，至少在一个相同的特征域下的域数据相同。所以，某个训练样本与第一训练样本相比，在任何相同的特征域下的域数据都不相同时，则不将这个训练样本作为第二训练样本。所以，上述的第二训练样本可能是多个训练样本中除第一训练样本之外的部分。

703：将第一训练样本和多个目标训练样本输入到特征信息提取网络，得到第一训练样本的目标特征信息。

示例性的，与上述图 4 示出的推荐方法类似，将第一训练样本和每个目标训练样本均作为输入数据输入到特征信息提取网络，得到第一训练样本的目标特征信息.

可选的，目标特征信息包括第四目标特征向量组（包括多个第四目标特征向量）和第五目标特征向量组（包括多个第五目标特征向量），其中，获取多个第四目标特征向量与上述获取多个第一目标特征向量的方式类似，即对第一训练样本的特征向量进行映射，得到多个第四目标特征向量，其中，第一训练样本的特征向量是对第一训练样本的每个用户特征域数据和每个物品特征域数据均进行编码得到，不再详细描述；可选的，获取多个第五目标特征向量与上述获取多个第二目标特征向量的方式类似，即对多个目标训练样本在同一个特征域下的多个第一特征向量进行融合得到，其中，每个目标训练样本对应的多个第一特征向量是对每个目标训练样本的特征向量进行映射得到，每个目标训练样本的特征向量为对每个目标训练样本的每个用户特征域数据和每个物品特征域数据均进行编码得到，也不再详细描述；

与上述获取第二特征向量组的方式类似，在获取第五特征向量组时，可以不对目标参考样本的标签数据进行向量化以及融合，也可以选择对目标参考样本的标签数据进行向量化以及融合。

可选的，目标特征信息还可以包括第六目标特征向量组（包括多个第六目标特征向量），其中，多个第六目标特征向量的获取方式与上述多个第三目标特征向量的获取方式类似，即将多个第四目标特征向量和多个第五目标特征向量进行拼接，得到第二向量组；然后，对第二向量组中的目标特征向量进行两两交互，得到多个第六目标特征向量，不再详细描述。

704：将目标特征信息输入到深度神经网络 DNN，得到输出值，输出值用于表征第一参考用户对第一参考物品进行操作的概率。

将第一训练样本的目标特征信息输入到推荐模型的多层感知器，得到输出值，即预测出第一参考用户对第一参考物品的操作情况。

705：根据输出值以及第一训练样本的标签数据进行推荐模型的训练，获得目标推荐模型。

示例性的，根据输出值和第一训练样本的标签数据确定损失，即根据预测出的第一参考用户对第一参考物品进行操作的情况和第一参考用户对第一参考物品进行操作的实际情况，确定损失；根据该损失以及梯度下降法调整待训练的推荐模型的模型参数，对推荐练模型进行训练，得到目标推荐模型。

应理解，对推荐模型的训练是使用多个训练样本进行迭代训练，其中，每个训练样本的训练过程与图 7 示出的使用第一训练样本的训练过程类似，不再叙述；直至推荐模型收敛时，完成对推荐模型的训练，得到目标推荐模型。

因此，基于上述的模型训练方法以及模型的应用过程，本申请的物品推荐流程和现有的物品推荐存在如图 8 所示的区别。如图 8 所示，现有的物品推荐过程是先使用训练样本进行

模型训练（与现有的有监督训练方法一致），待完成模型训练之后，将待预测数据直接输入到推荐模型中进行用户行为预测，得到输出值，基于输出值确定是否向用户进行物品推荐；而本申请的用户行为预测是先使用训练样本对待推荐模型进行训练（与图 7 示出的训练方法一致），待完成模型训练之后，将训练样本作为参考样本；当获取到待预测数据时，先从训练样本中获取与待预测数据对应的目标训练样本，然后，将待预测数据和目标训练样本一起输入到目标推荐模型中进行用户行为预测，获得输出值，基于该输出值确定是否向用户进行物品推荐。由于本申请的用户行为预测过程中会融合目标训练样本的特征信息，得到丰富的目标特征信息，提高用户行为预测的精度，提高输出值的精度，从而可以进行精确的物品推荐。

最后结合附图具体介绍一下应用本申请的推荐方法后的几种常见推荐场景。

应用 1：针对点击概率的预测。

如图 9 所示，针对应用程序的推荐，比如，精品应用下的应用程序推荐、精品新游榜单下的应用程序推荐；针对每种类型的应用程序推荐，首先获取多个候选应用程序；然后，基于目标用户的用户特征域数据以及各个候选应用程序的物品特征域数据，构造与各个候选应用程序对应的待预测数据，即将目标用户的用户特征域数据和各个候选应用程序的物品特征域数据拼接为一个待预测数据；然后，基于上述的推荐方法以及待预测数据，预测出目标用户对各个候选应用程序的点击概率；然后，根据各个候选应用的点击概率从高到低的顺序对多个候选应用程序进行排序，并在推荐页面按照点击概率从高到低的顺序展示排序后的多个候选应用程序或者只展示排序靠前的候选应用程序。

由于本申请中在进行点击率预测时，会融合目标参考样本的特征信息，从而使预测出的点击率更加准确，从而向目标用户推荐的应用程序更加精确，进而提高应用程序的下载率。

应用 2：针对购买概率的预测。

如图 10 所示，针对商品的推荐，可基于各个候选商品的物品特征域数据以及目标用户的用户特征域数据，构造与各个候选商品对应的待预测数据，即将目标用户的用户特征域数据和各个候选商品的物品特征域数据拼接为一个待预测数据；基于上述的推荐方法以及待预测数据，预测出目标用户对每个候选商品的购买概率；然后，根据每个候选商品的购买概率从高到低的顺序对多个候选商品进行排序，并在推荐页面按照购买概率从高到低的顺序展示排序后的多个候选商品或者只展示排序靠前的候选商品。

由于本申请中在进行购买概率预测时，会融合目标参考样本的特征信息，从而使预测出的购买概率更加准确，从而向目标用户推荐的商品更加符合用户的需求，提高商品的销量。

应用 3：针对歌曲的评分的预测。

如图 11 所示，针对歌曲推荐，比如，私人 FM 中的歌曲推荐、每日 30 首中的歌曲推荐；针对每种类型下的推荐，首先获取多个候选歌曲，根据各个候选歌曲的物品特征域数据以及目标用户的用户特征域数据，构造与每个候选歌曲对应的待预测数据，即将目标用户的用户特征域数据和各个候选歌曲的物品特征域数据拼接为一个待预测数据；基于上述的推荐方法以及待预测数据，预测出每个候选歌曲的评分，每个候选歌曲的评分用于表征目标用户对该候选歌曲的喜爱程度；然后，根据每个候选歌曲的评分从高到低的顺序对多个候选歌曲进行排序，并在推荐页面按照评分从高到低的顺序展示排序后的多个候选歌曲或者只展示评分靠前的候选歌曲。

由于本申请中在对歌曲进行评分预测时，会融合目标参考样本的特征信息，从而使预测出的评分更加准确，从而推荐的歌曲更加符合用户的需求，提高歌曲推荐的准确率。

将本申请的用户行为建模方法与现有的基于特征交互的用户行为建模方法以及基于用户

行为序列的用户行为建模方法相比，在 CTR 预估任务上进行了充分的实验，实验设置如下：

实验设置 1：使用以下测试指标评估模型的预测精度的优劣，即：

接受者操作特性曲线下与坐标轴围成的面积（Area Under Curve，AUC）、损失（Logloss，LL）以及相对提升（relative improvement，REI.Impr）；其中，对于测试指标 AUC 来说，取值越大说明模型的效果越好；对于测试指标 LL 来说，取值越小说明模型效果越好；其中，测试 REI.Impr 是本申请的模型（RIM）相对于其他模型的预测精度的提升，则对于测试指标 REI.Impr 来说，取值越大说明 RIM 的预测精度相对于被比较的模型的精度越高。

实验设置 2：获取应用 A 的数据集、应用 B 的数据集以及应用 C 的数据集，分别在应用 A 的数据集、应用 B 的数据集以及应用 C 的数据集上测试基于用户行为预测 CTR 的模型的 AUC 和 LL，以及测试本申请的模型预测 CTR 时的 AUC 和 LL；

实验设置 3：获取第一数据集和第二数据集，分别在第一数据集和第二数据集上测试基于特征交互预测 CTR 的模型的 AUC 和 LL，以及测试本申请的模型预测 CTR 时的 AUC 和 LL。例如，第一数据集可以为 avazu，第二数据集可以为 criteo。

其中，基于用户行为预测 CTR 的模型包括：HPMN，MIMN，DIN，DIEN，SIM，UBR；基于特征交互预测 CTR 的模型包括：LR，GBDT，FM，FFM，AFM，FNN，DeepFM，IPNN，PIN，xDeepFM，FGCNN。

其中，表 3 和表 4 为对比结果。

表 3：与基于用户行为预测 CTR 的模型的对比结果：

Model	Tmall			Taobao			Alipay		
	AUC	LL	RelImpr	AUC	LL	RelImpr	AUC	LL	RelImpr
HPMN	0.8526	0.4976	7.17%	0.7599	0.5911	12.68%	0.7681	0.5976	4.23%
MIMN	0.8457	0.5008	8.05%	0.7533	0.6002	13.67%	0.7667	0.5998	4.42%
DIN	0.8796	0.4292	3.88%	0.7433	0.6086	15.20%	0.7647	0.6044	4.69%
DIEN	0.8838	0.4445	3.39%	0.7506	0.6084	14.08%	0.7502	0.6151	6.71%
SIM	0.8857	0.4520	3.17%	0.7825	0.5795	9.43%	0.7600	0.6089	5.34%
UBR	0.8975	0.4368	1.81%	0.8169	0.5432	4.82%	0.7952	0.5747	0.67%
RIM	0.9138	0.3804	0.00%	0.8563	0.4644	0.00%	0.8006	0.5615	0.00%

表 4：与基于特征交互预测 CTR 的模型的对比结果：

Model	Avazu			Criteo		
	AUC	LL	RelImpr	AUC	LL	RelImpr
LR	76.76%	0.3868	2.85%	78.00%	0.5631	2.88%
GBDT	77.53%	0.3824	1.83%	78.62%	0.5560	2.07%
FM	77.93%	0.3805	1.31%	79.09%	0.5500	1.47%
FFM	78.31%	0.3781	0.82%	79.80%	0.5438	0.56%
AFM	78.06%	0.3794	1.14%	79.13%	0.5517	1.42%
FNN	78.30%	0.3778	0.83%	79.87%	0.5428	0.48%
DeepFM	78.36%	0.3777	0.75%	79.91%	0.5423	0.43%
IPNN	78.68%	0.3757	0.34%	80.13%	0.5399	0.15%
PIN	78.72%	0.3755	0.29%	80.18%	0.5394	0.09%
xDeepFM	78.55%	0.3766	0.51%	80.06%	0.5408	0.24%
FGCNN	78.82%	0.3747	0.16%	80.22%	0.5389	0.04%
RIM	78.93%	0.3741	0.00%	80.25%	0.5387	0.00%

经过实验可以看出在预测精度上，本申请的模型在 AUC 和 Logloss 两个指标上都取得了最好的实验效果。因此，使用本申请的模型预测用户行为，会使预测结果更加精确，可以获得更加精确的输出值，从而为用户进行更精确的推荐。

参阅图 12，图 12 为本申请实施例提供的一种推荐装置的结构图。推荐装置 1200 包括获

取单元 1201 和处理单元 1202；

获取单元 1201，用于获取待预测数据；

处理单元 1202，用于获取待预测数据；根据待预测数据和多个参考样本的相似度从多个参考样本中获取多个目标参考样本；每个参考样本和待预测数据均包括用户特征域数据和物品特征域数据，待预测数据的用户特征域数据用于指示目标用户特征，待预测数据的物品特征域数据用于指示目标物品特征，每个目标参考样本和待预测数据具有部分相同的用户特征域数据和/或物品特征域数据；根据多个目标参考样本与待预测数据获取待预测数据的目标特征信息；目标特征信息包括第一目标特征向量组和第二目标特征向量组，第一目标特征向量组为向量化后的待预测数据，第二目标特征向量组为对多个目标参考样本进行向量化后融合得到；以目标特征信息为输入通过深度神经网络 DNN 获取输出值；根据输出值确定是否向目标用户推荐目标物品。

关于上述获取单元 1201 和处理单元 1202 更详细的描述，可参考上述方法实施例中的相关描述，在此不再说明。

参阅图 13，图 13 为本申请实施例提供的一种推荐模型训练装置的结构图。推荐模型包括特征信息提取网络和深度神经网络 DNN。推荐模型训练装置 1300 包括获取单元 1301 和处理单元 1302；

获取单元 1301，用于获取多个训练样本，其中，每个训练样本包括用户特征域数据和物品特征域数据；

处理单元 1302，用于根据第一训练样本和多个第二训练样本的相似度从多个第二训练样本中获取多个目标训练样本，其中，第一训练样本为多个训练样本中的一个，多个第二训练样本为多个训练样本除第一训练样本之外的部分或全部，第一训练样本的用户特征域数据用于指示第一参考用户特征，第一训练样本的物品特征域数据用于指示第一参考物品特征，第一训练样本和每个目标训练样本具有部分相同的用户特征域数据和/或物品特征域数据；

将第一训练样本和多个目标训练样本输入到特征信息提取网络，得到第一训练样本的目标特征信息，其中，目标特征信息包括第四目标特征向量组和第五目标特征向量组，第四目标特征向量组为通过特征信息提取网络对第一训练样本进行向量化得到，第五目标特征向量组为通过特征信息提取网络对多个目标训练样本进行向量化后融合得到；

将目标特征信息输入到深度神经网络 DNN，得到输出值，输出值用于表征第一参考用户对第一参考物品进行操作的概率；

根据输出值以及第一训练样本的标签进行推荐模型的训练，获得目标推荐模型。

关于上述获取单元 1301 和处理单元 1302 更详细的描述，可参考上述方法实施例中的相关描述，在此不再说明。

参阅图 14，图 14 为本申请实施例提供的一种电子设备的结构图。电子设备 1400 可以为上述的推荐装置 1200；或者，为推荐装置 1200 中的芯片或芯片系统；电子设备还可以为上述的推荐模型训练装置 1300；或者，为推荐模型训练装置 1300 中的芯片或芯片系统；

电子设备 1400 包括存储器 1401、处理器 1402、通信接口 1403 以及总线 1404。其中，存储器 1401、处理器 1402、通信接口 1403 通过总线 1404 实现彼此之间的通信连接。

存储器 1401 可以是只读存储器（Read Only Memory，ROM），静态存储设备，动态存储设备或者随机存取存储器（Random Access Memory，RAM）。存储器 1401 可以存储程序，当存储器 1401 中存储的程序被处理器 1402 执行时，处理器 1402 和通信接口 1403 用于执行本申请实施例的数据流传输方法中的各个步骤。

处理器 1402 可以采用通用的中央处理器（Central Processing Unit，CPU），微处理器，应用专用集成电路（Application Specific Integrated Circuit，ASIC），图形处理器（graphics processing unit，GPU）或者一个或多个集成电路，用于执行相关程序，以实现本申请实施例的音频特征补偿装置或音频识别装置中的单元所需执行的功能，或者执行本申请方法实施例的数据流传输方法。

处理器 1402 还可以是一种集成电路芯片，具有信号的处理能力。在实现过程中，本申请的数据流传输方法中的各个步骤可以通过处理器 1402 中的硬件的集成逻辑电路或者软件形式的指令完成。上述的处理器 1402 还可以是通用处理器、数字信号处理器（Digital Signal Processing，DSP）、专用集成电路（ASIC）、现成可编程门阵列（Field Programmable Gate Array，FPGA）或者其他可编程逻辑器件、分立门或者晶体管逻辑器件、分立硬件组件。可以实现或者执行本申请实施例中的公开的各方法、步骤及逻辑框图。通用处理器可以是微处理器或者该处理器也可以是任何常规的处理器等。结合本申请实施例所公开的方法的步骤可以直接体现为硬件译码处理器执行完成，或者用译码处理器中的硬件及软件模块组合执行完成。软件模块可以位于随机存储器，闪存、只读存储器，可编程只读存储器或者电可擦写可编程存储器、寄存器等本领域成熟的存储介质中。该存储介质位于存储器 1401，处理器 1402 读取存储器 1401 中的信息，结合其硬件完成本申请实施例的用户设备或头戴设备中包括的单元所需执行的功能，或者执行本申请方法实施例的数据流传输方法中的各个步骤。

通信接口 1403 可以为收发器一类的收发装置，来实现电子设备 1400 与其他设备或通信网络之间的通信；通信接口 1403 也可以为输入-输出接口，来实现电子设备 1400 与输入-输出设备之间的数据传输，其中，输入-输出设备包括但不限于键盘、鼠标、显示屏、U 盘以及硬盘。比如，处理器 1402 可以通过通信接口 1403 获取待预测数据。

总线 1404 可包括在装置电子设备 1400 各个部件（例如，存储器 1401、处理器 1402、通信接口 1403）之间传送信息的通路。

应注意，尽管图 14 所示电子设备 1400 仅仅示出了存储器、处理器、通信接口，但是在具体实现过程中，本领域的技术人员应当理解，电子设备 1400 还包括实现正常运行所必须的其他器件。同时，根据具体需要，本领域的技术人员应当理解，电子设备 1400 还可包括实现其他附加功能的硬件器件。此外，本领域的技术人员应当理解，电子设备 1400 也可仅仅包括实现本申请实施例所必须的器件，而不必包括图 14 中所示的全部器件。

在本申请所提供的几个实施例中，应该理解到，所揭露的系统、装置和方法，可以通过其它的方式实现。例如，以上所描述的装置实施例仅仅是示意性的，例如，所述单元的划分，仅仅为一种逻辑功能划分，实际实现时可以有另外的划分方式，例如多个单元或组件可以结合或者可以集成到另一个系统，或一些特征可以忽略，或不执行。另一点，所显示或讨论的相互之间的耦合或直接耦合或通信连接可以是通过一些接口，装置或单元的间接耦合或通信连接，可以是电性，机械或其它的形式。

所述作为分离部件说明的单元可以是或者也可以不是物理上分开的，作为单元显示的部件可以是或者也可以不是物理单元，即可以位于一个地方，或者也可以分布到多个网络单元上。可以根据实际的需要选择其中的部分或者全部单元来实现本实施例方案的目的。

另外，在本申请各个实施例中的各功能单元可以集成在一个处理单元中，也可以是各个单元单独物理存在，也可以两个或两个以上单元集成在一个单元中。

本申请中，“至少一个”是指一个或者多个，“多个”是指两个或两个以上。“和/或”，描述关联对象的关联关系，表示可以存在三种关系，例如，A 和/或 B，可以表示：单独存在 A，

同时存在A和B，单独存在B的情况，其中A，B可以是单数或者复数。在本申请的文字描述中，字符“/”，一般表示前后关联对象是一种“或”的关系；在本申请的公式中，字符“/”，表示前后关联对象是一种“相除”的关系。

可以理解的是，在本申请的实施例中涉及的各种数字编号仅为描述方便进行的区分，并不用来限制本申请的实施例的范围。上述各过程的序号的大小并不意味着执行顺序的先后，各过程的执行顺序应以其功能和内在逻辑确定。

所述功能如果以软件功能单元的形式实现并作为独立的产品销售或使用时，可以存储在一个计算机可读取存储介质中。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或者该技术方案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中，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）执行本申请各个实施例所述方法的全部或部分步骤。而前述的存储介质包括：U盘、移动硬盘、只读存储器（Read-Only Memory，ROM）、随机存取存储器（Random Access Memory，RAM）、磁碟或者光盘等各种可以存储程序代码的介质。

以上所述，仅为本申请的具体实施方式，但本申请的保护范围并不局限于此，任何熟悉本技术领域的技术人员在本申请揭露的技术范围内，可轻易想到变化或替换，都应涵盖在本申请的保护范围之内。因此，本申请的保护范围应以所述权利要求的保护范围为准。

权 利 要 求 书

1.一种推荐方法，其特征在于，包括：

获取待预测数据；

根据所述待预测数据和多个参考样本的相似度从所述多个参考样本中获取多个目标参考样本；每个所述参考样本和所述待预测数据均包括用户特征域数据和物品特征域数据，所述待预测数据的所述用户特征域数据用于指示目标用户特征，所述待预测数据的所述物品特征域数据用于指示目标物品特征，每个所述目标参考样本和所述待预测数据具有部分相同的用户特征域数据和/或物品特征域数据；

根据所述多个目标参考样本与所述待预测数据获取所述待预测数据的目标特征信息；所述目标特征信息包括第一目标特征向量组和第二目标特征向量组，所述第一目标特征向量组为向量化后的所述待预测数据，所述第二目标特征向量组为对所述多个目标参考样本进行向量化后融合得到；

以所述目标特征信息为输入通过深度神经网络DNN获取输出值；

根据所述输出值确定是否向所述目标用户推荐所述目标物品。

2.根据权利要求1所述的方法，其特征在于，

所述多个目标参考样本还包括标签数据；

所述第二目标特征向量组为对所述多个目标参考样本进行向量化后融合得到，具体为：

所述第二目标特征向量组为对所述多个目标参考样本的用户特征域数据，物品特征域数据以及标签数据进行向量化后融合得到。

3.根据权利要求1或2所述的方法，其特征在于，

所述目标特征信息还包括第三目标特征向量组，所述第三目标特征向量组为对第一向量组中的目标特征向量进行两两交互得到，所述第一向量组包括所述第一目标特征向量组和所述第二目标特征向量组。

4.根据权利要求1-3中任一项所述的方法，其特征在于，所述融合包括：

对所述第一目标特征向量组中的多个第一目标特征向量进行拼接，得到所述待预测数据的第二特征向量；

对每个所述目标参考样本的多个第一特征向量进行拼接，得到每个所述目标参考样本的第二特征向量，每个所述目标参考样本的多个第一特征向量为对所述目标参考样本进行向量化得到；

获取每个所述目标参考样本的第二特征向量与所述待预测数据的第二特征向量之间的相似度；

根据每个所述目标参考样本的第二特征向量与所述待预测数据的第二特征向量之间的相似度，确定每个所述目标参考样本的权重；

根据每个所述目标参考样本的权重，对所述多个目标参考样本在同一个特征域下的第一特征向量进行融合，得到所述第二目标特征向量组。

5.根据权利要求1-4中任一项所述的方法，其特征在于，根据所述待预测数据和多个参考样本的相似度从所述多个参考样本中获取多个目标参考样本之前，所述方法还包括：

获取多个原始样本，其中，每个所述原始样本包括用户特征域数据和物品特征域数据；

将所述待预测数据的多个所述用户特征域数据和多个所述物品特征域数据作为元素，对所述多个原始样本进行倒排索引，得到所述多个参考样本。

6.一种推荐模型训练方法，其特征在于，所述推荐模型包括特征信息提取网络和深度神经网络DNN，所述方法包括：

获取多个训练样本，其中，每个所述训练样本包括用户特征域数据和物品特征域数据；

根据第一训练样本和多个第二训练样本的相似度从所述多个第二训练样本中获取多个目标训练样本，其中，所述第一训练样本为所述多个训练样本中的一个，所述多个第二训练样本为所述多个训练样本除所述第一训练样本之外的部分或全部，所述第一训练样本的所述用户特征域数据用于指示第一参考用户特征，所述第一训练样本的所述物品特征域数据用于指示第一参考物品特征，所述第一训练样本和每个所述目标训练样本具有部分相同的用户特征域数据和/或物品特征域数据；

将所述第一训练样本和所述多个目标训练样本输入到所述特征信息提取网络，得到所述第一训练样本的目标特征信息，其中，所述目标特征信息包括第四目标特征向量组和第五目标特征向量组，所述第四目标特征向量组为通过所述特征信息提取网络对所述第一训练样本进行向量化得到，所述第五目标特征向量组为通过所述特征信息提取网络对所述多个目标训练样本进行向量化后融合得到；

将所述目标特征信息输入到所述深度神经网络DNN，得到输出值，所述输出值用于表征所述第一参考用户对所述第一参考物品进行操作的概率；

根据所述输出值以及所述第一训练样本的标签数据进行所述推荐模型的训练，获得目标推荐模型。

7.根据权利要求6所述的方法，其特征在于，

所述第五目标特征向量组为对所述多个目标训练样本进行向量化后融合得到，具体为：

所述第五目标特征向量组为通过所述特征信息提取网络对所述多个目标训练样本的用户特征域数据、物品特征域数据以及标签数据进行向量化后融合得到。

8.根据权利要求6或7所述的方法，其特征在于，

所述目标特征信息还包括第六目标特征向量组，所述第六目标特征向量组是通过所述特征信息提取网络对第二向量组中的目标特征向量进行两两交互得到，所述第二向量组包括所述第四目标特征向量组和所述第五目标特征向量组。

9.根据权利要求6-8中任一项所述的方法，其特征在于，所述融合包括：

对所述第四目标特征向量组中的多个第四目标特征向量进行拼接，得到所述第一训练样本的第二特征向量；

对每个所述目标训练样本的多个第一特征向量进行拼接，得到每个所述目标训练样本的第二特征向量，每个所述目标训练样本的多个第一特征向量为对所述目标训练样本进行向量化得到；

获取每个所述目标训练样本的第二特征向量与所述第一训练样本的第二特征向量之间的相似度；

根据每个所述目标训练样本的第二特征向量与所述第一训练样本的第二特征向量之间的相似度，确定每个所述目标训练样本的权重；

根据每个所述目标训练样本的权重，对所述多个目标训练样本在同一个特征域下的第一特征向量进行融合，得到所述第五目标特征向量组。

10.根据权利要求 6-9 中任一项所述的方法，其特征在于，根据第一训练样本和多个第二训练样本的相似度从所述多个第二训练样本中获取多个目标训练样本之前，所述方法还包括：

将所述第一训练样本的多个所述用户特征域数据和多个所述物品特征域数据作为元素，对所述多个训练样本进行倒排索引，得到所述多个第二训练样本。

11.一种推荐装置，其特征在于，包括：获取单元和处理单元；

所述获取单元，用于获取待预测数据；

所述处理单元，用于根据所述待预测数据和多个参考样本的相似度从所述多个参考样本中获取多个目标参考样本；每个所述参考样本和所述待预测数据均包括用户特征域数据和物品特征域数据，所述待预测数据的所述用户特征域数据用于指示目标用户特征，所述待预测数据的所述物品特征域数据用于指示目标物品特征，每个所述目标参考样本和所述待预测数据具有部分相同的用户特征域数据和/或物品特征域数据；

根据所述多个目标参考样本与所述待预测数据获取所述待预测数据的目标特征信息；所述目标特征信息包括第一目标特征向量组和第二目标特征向量组，所述第一目标特征向量组为向量化后的所述待预测数据，所述第二目标特征向量组为对所述多个目标参考样本进行向量化后融合得到；

以所述目标特征信息为输入通过深度神经网络 DNN 获取输出值；

根据所述输出值确定是否向所述目标用户推荐所述目标物品。

12.根据权利要求 11 所述的装置，其特征在于，

所述多个目标参考样本还包括标签数据；

所述第二目标特征向量组为对所述多个目标参考样本进行向量化后融合得到，具体为：

所述第二目标特征向量组为对所述多个目标参考样本的用户特征域数据，物品特征域数据以及标签数据进行向量化后融合得到。

13.根据权利要求 12 所述的装置，其特征在于，

所述目标特征信息还包括第三目标特征向量组，所述第三目标特征向量组为对第一向量组中的目标特征向量进行两两交互得到，所述第一向量组包括所述第一目标特征向量组和所述第二目标特征向量组。

14.根据权利要求 11-13 中任一项所述的装置，其特征在于，

在所述处理单元进行融合方面，所述处理单元，具体用于：

对所述第一目标特征向量组中的多个第一目标特征向量进行拼接，得到所述待预测数据的第二特征向量；

对每个所述目标参考样本的多个第一特征向量进行拼接，得到每个所述目标参考样本的第二特征向量，每个所述目标参考样本的多个第一特征向量为对所述目标参考样本进行向量

化得到；

获取每个所述目标参考样本的第二特征向量与所述待预测数据的第二特征向量之间的相似度；

根据每个所述目标参考样本的第二特征向量与所述待预测数据的第二特征向量之间的相似度，确定每个所述目标参考样本的权重；

根据每个所述目标参考样本的权重，对所述多个目标参考样本在同一个特征域下的第一特征向量进行融合，得到所述第二目标特征向量组。

15.根据权利要求 11-14 中任一项所述的装置，其特征在于，

在所述处理单元根据所述待预测数据和多个参考样本的相似度从所述多个参考样本中获取多个目标参考样本之前，所述处理单元，还用于获取多个原始样本，其中，每个所述原始样本包括用户特征域数据和物品特征域数据；

将所述待预测数据的多个所述用户特征域数据和多个所述物品特征域数据作为元素，对所述多个原始样本进行倒排索引，得到所述多个参考样本。

16.一种推荐模型训练装置，其特征在于，所述推荐模型包括特征信息提取网络和深度神经网络 DNN，所述装置包括：获取单元和处理单元；

所述获取单元，用于获取多个训练样本，其中，每个所述训练样本包括用户特征域数据和物品特征域数据；

所述处理单元，用于根据第一训练样本和多个第二训练样本的相似度从所述多个第二训练样本中获取多个目标训练样本，其中，所述第一训练样本为所述多个训练样本中的一个，所述多个第二训练样本为所述多个训练样本除所述第一训练样本之外的部分或全部，所述第一训练样本的所述用户特征域数据用于指示第一参考用户特征，所述第一训练样本的所述物品特征域数据用于指示第一参考物品特征，所述第一训练样本和每个所述目标训练样本具有部分相同的用户特征域数据和/或物品特征域数据；

将所述第一训练样本和所述多个目标训练样本输入到所述特征信息提取网络，得到所述第一训练样本的目标特征信息，其中，所述目标特征信息包括第四目标特征向量组和第五目标特征向量组，所述第四目标特征向量组为通过所述特征信息提取网络对所述第一训练样本进行向量化得到，所述第五目标特征向量组为通过所述特征信息提取网络对所述多个目标训练样本进行向量化后融合得到；

将所述目标特征信息输入到所述深度神经网络 DNN，得到输出值，所述输出值用于表征所述第一参考用户对所述第一参考物品进行操作的概率；

根据所述输出值以及所述第一训练样本的标签进行所述推荐模型的训练，获得目标推荐模型。

17.根据权利要求 16 所述的装置，其特征在于，

所述第五目标特征向量组为对所述多个目标训练样本进行向量化后融合得到，具体为：

所述第五目标特征向量组为通过所述特征信息提取网络对所述多个目标训练样本的用户特征域数据、物品特征域数据以及标签数据进行向量化后融合得到。

18.根据权利要求 16 或 17 所述的装置，其特征在于，

所述目标特征信息还包括第六目标特征向量组，所述第六目标特征向量组是通过所述特征信息提取网络对第二向量组中的目标特征向量进行两两交互得到，所述第二向量组包括所述第四目标特征向量组和所述第五目标特征向量组。

19.根据权利要求 16-18 中任一项所述的装置，其特征在于，

在所述处理单元进行融合方面，所述处理单元，具体用于：

对所述第四目标特征向量组中的多个第四目标特征向量进行拼接，得到所述第一训练样本的第二特征向量；

对每个所述目标训练样本的多个第一特征向量进行拼接，得到每个所述目标训练样本的第二特征向量，每个所述目标训练样本的多个第一特征向量为对所述目标训练样本进行向量化得到；

获取每个所述目标训练样本的第二特征向量与所述第一训练样本的第二特征向量之间的相似度；

根据每个所述目标训练样本的第二特征向量与所述第一训练样本的第二特征向量之间的相似度，确定每个所述目标训练样本的权重；

根据每个所述目标训练样本的权重，对所述多个目标训练样本在同一个特征域下的第一特征向量进行融合，得到所述第五目标特征向量组。

20.根据权利要求 16-19 中任一项所述的装置，其特征在于，

在所述处理单元根据第一训练样本和多个第二训练样本的相似度从所述多个第二训练样本中获取多个目标训练样本之前，所述处理单元，将所述第一训练样本的多个所述用户特征域数据和多个所述物品特征域数据作为元素，对所述多个训练样本进行倒排索引，得到所述多个第二训练样本。

21.一种电子设备，其特征在于，包括：存储器，用于存储程序；处理器，用于执行存储器存储的程序；当存储器存储的程序被执行时，处理器用于实现权利要求 1-5 或权利要求 6-10 中任一项所述的方法。

22.一种计算机可读存储介质，其特征在于，所述计算机可读存储介质存储用于设备执行的程序代码，所述程序代码包括用于实现权利要求 1-5 或权利要求 6-10 中任一项所述的方法。

23.一种计算程序产品，其特征在于，当所述计算程序产品在计算机上运行时，使得计算机实现权利要求 1-5 或权利要求 6-10 中任一项所述的方法。

24.一种芯片，其特征在于，所述芯片包括处理器与数据接口，所述处理器通过数据接口读取所述存储器上存储的指令，实现权利要求 1-5 或权利要求 6-10 中任一项所述的方法。

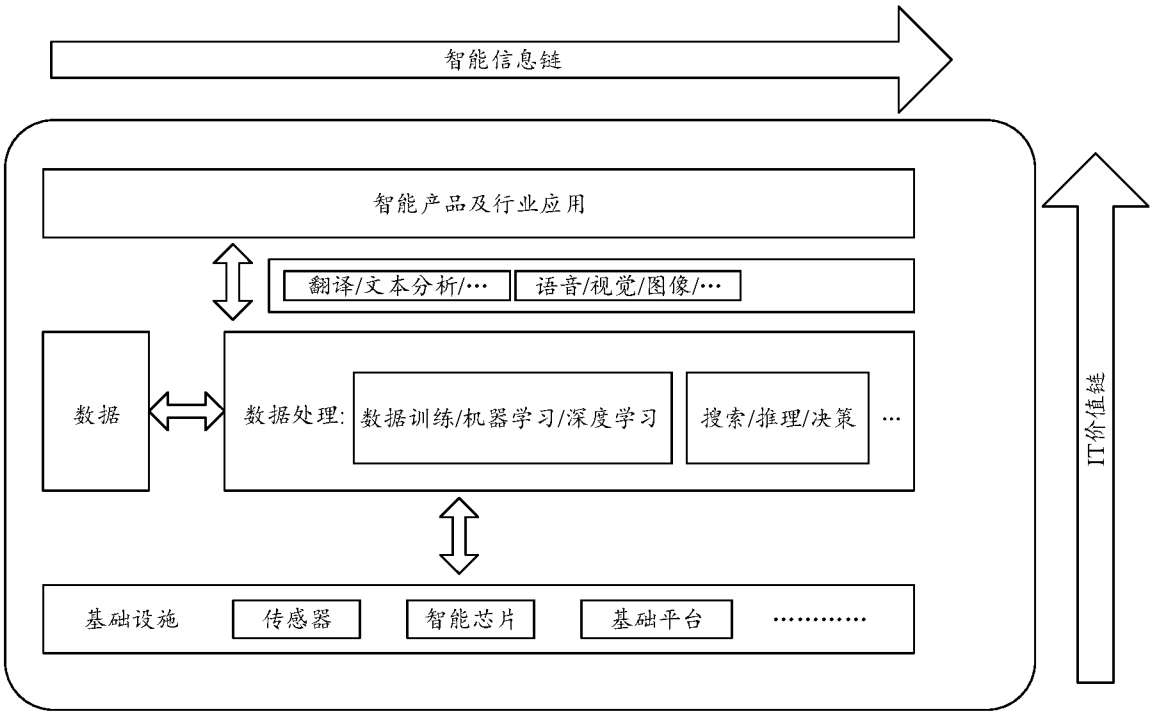

智能信息链
智能产品及行业应用
翻译/文本分析/…
语音/视觉/图像/…
数据
数据处理:
数据训练/机器学习/深度学习
搜索/推理/决策
…
基础设施
传感器
智能芯片
基础平台
……
IT价值链

图 1

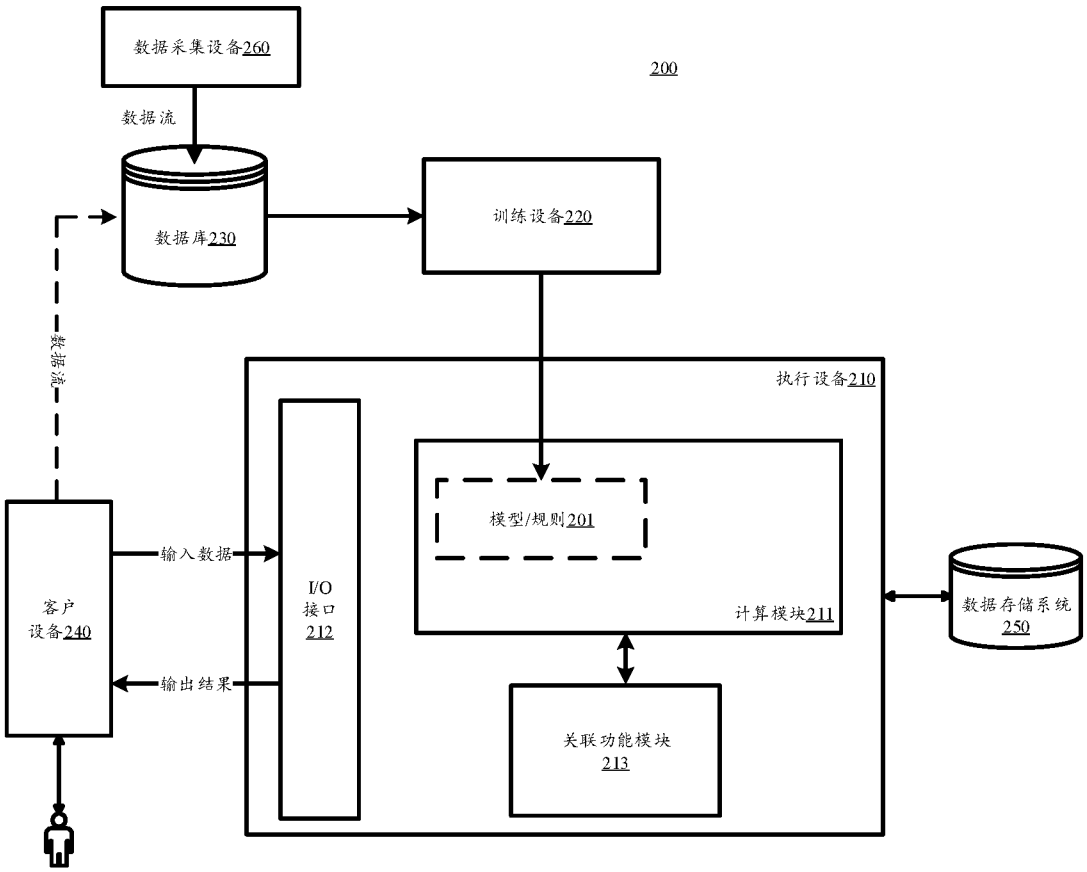

数据采集设备260
200
数据流
数据库230
训练设备220
数据流
执行设备210
模型/规则201
计算模块211
数据存储系统
250
客户
设备240
输入数据
输出结果
I/O
接口
212
关联功能模块
213

图 2

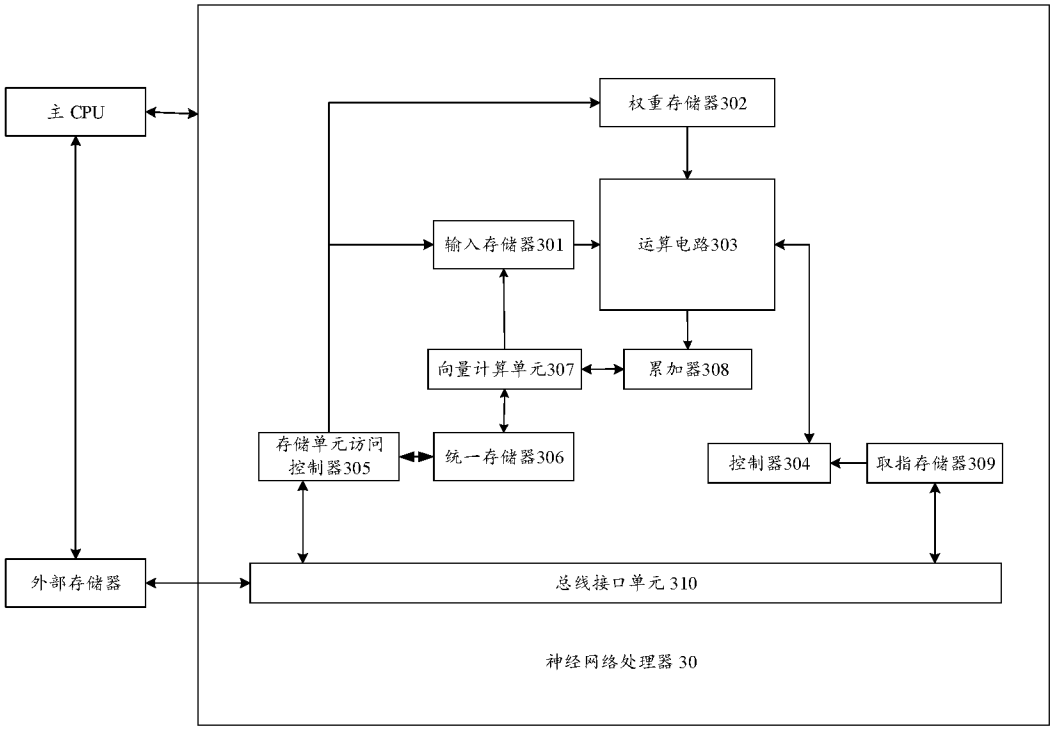

主 CPU
权重存储器302
输入存储器301
运算电路303
向量计算单元307
累加器308
存储单元访问控制器305
统一存储器306
控制器304
取指存储器309
外部存储器
总线接口单元 310
神经网络处理器 30

图 3

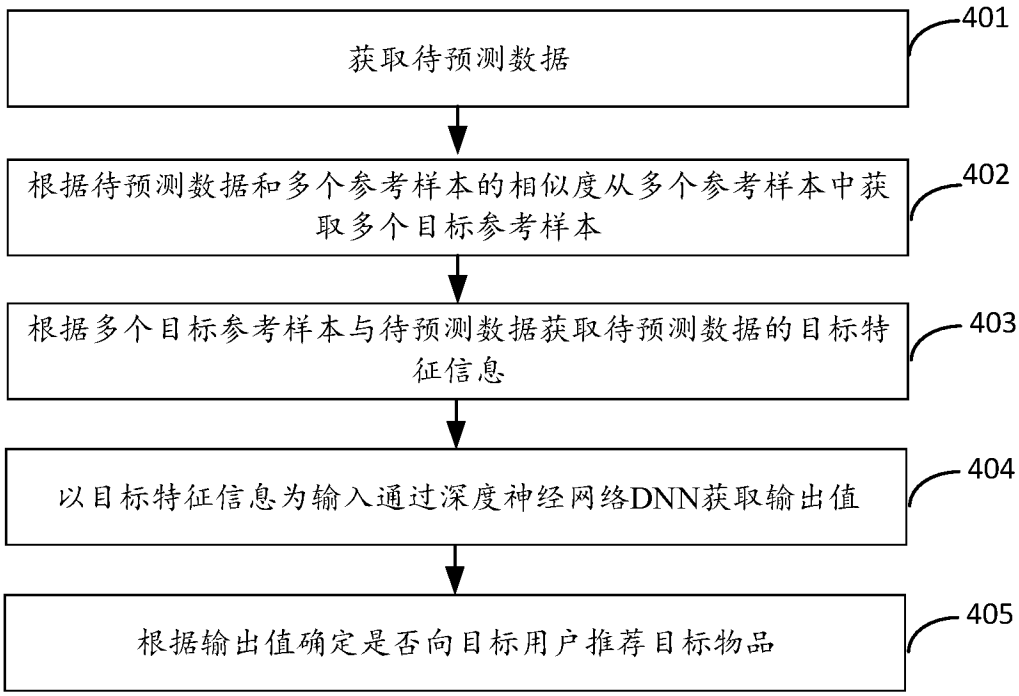

获取待预测数据
401
根据待预测数据和多个参考样本的相似度从多个参考样本中获取多个目标参考样本
402
根据多个目标参考样本与待预测数据获取待预测数据的目标特征信息
403
以目标特征信息为输入通过深度神经网络DNN获取输出值
404
根据输出值确定是否向目标用户推荐目标物品
405

图 4

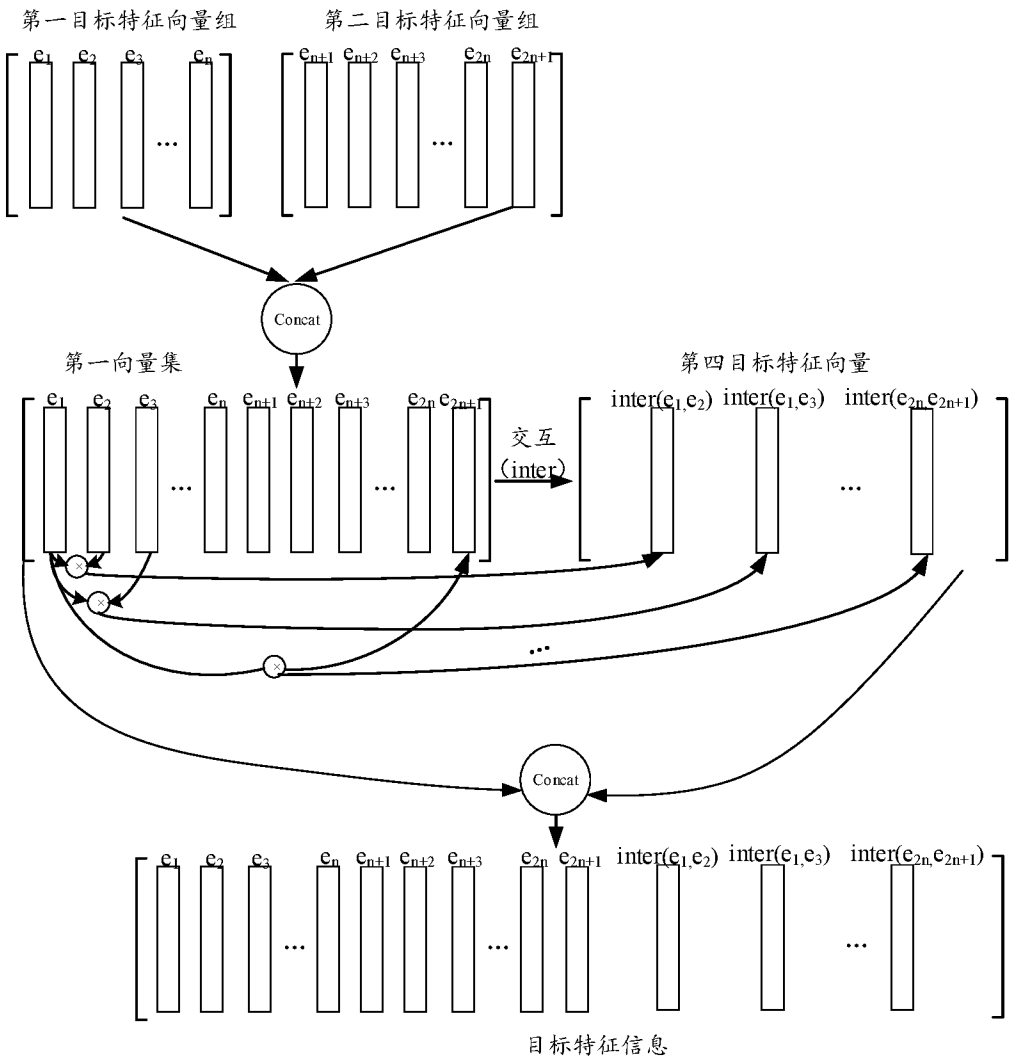

第一目标特征向量组
第二目标特征向量组
Concat
第一向量集
第四目标特征向量
交互
(inter)
Concat
目标特征信息

图 5

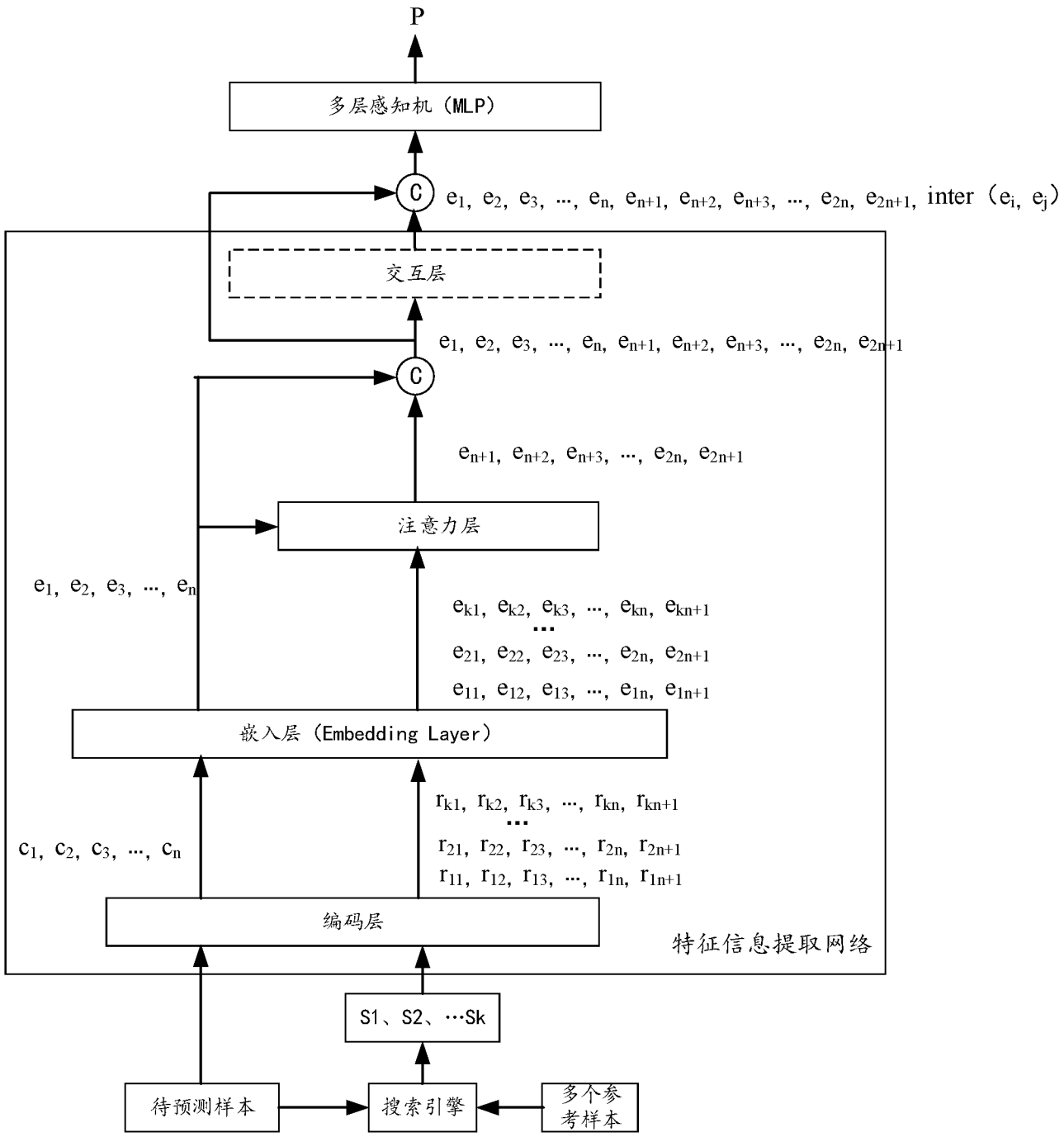

P
多层感知机（MLP）
C
$e_1, e_2, e_3, \cdots, e_n, e_{n+1}, e_{n+2}, e_{n+3}, \cdots, e_{2n}, e_{2n+1}$, inter（$e_i, e_j$）
交互层
$e_1, e_2, e_3, \cdots, e_n, e_{n+1}, e_{n+2}, e_{n+3}, \cdots, e_{2n}, e_{2n+1}$
C
$e_{n+1}, e_{n+2}, e_{n+3}, \cdots, e_{2n}, e_{2n+1}$
注意力层
$e_1, e_2, e_3, \cdots, e_n$
$e_{k1}, e_{k2}, e_{k3}, \cdots, e_{kn}, e_{kn+1}$
$e_{21}, e_{22}, e_{23}, \cdots, e_{2n}, e_{2n+1}$
$e_{11}, e_{12}, e_{13}, \cdots, e_{1n}, e_{1n+1}$
嵌入层（Embedding Layer）
$r_{k1}, r_{k2}, r_{k3}, \cdots, r_{kn}, r_{kn+1}$
$r_{21}, r_{22}, r_{23}, \cdots, r_{2n}, r_{2n+1}$
$r_{11}, r_{12}, r_{13}, \cdots, r_{1n}, r_{1n+1}$
$c_1, c_2, c_3, \cdots, c_n$
编码层
特征信息提取网络
S1、S2、…Sk
待预测样本
搜索引擎
多个参考样本

图 6

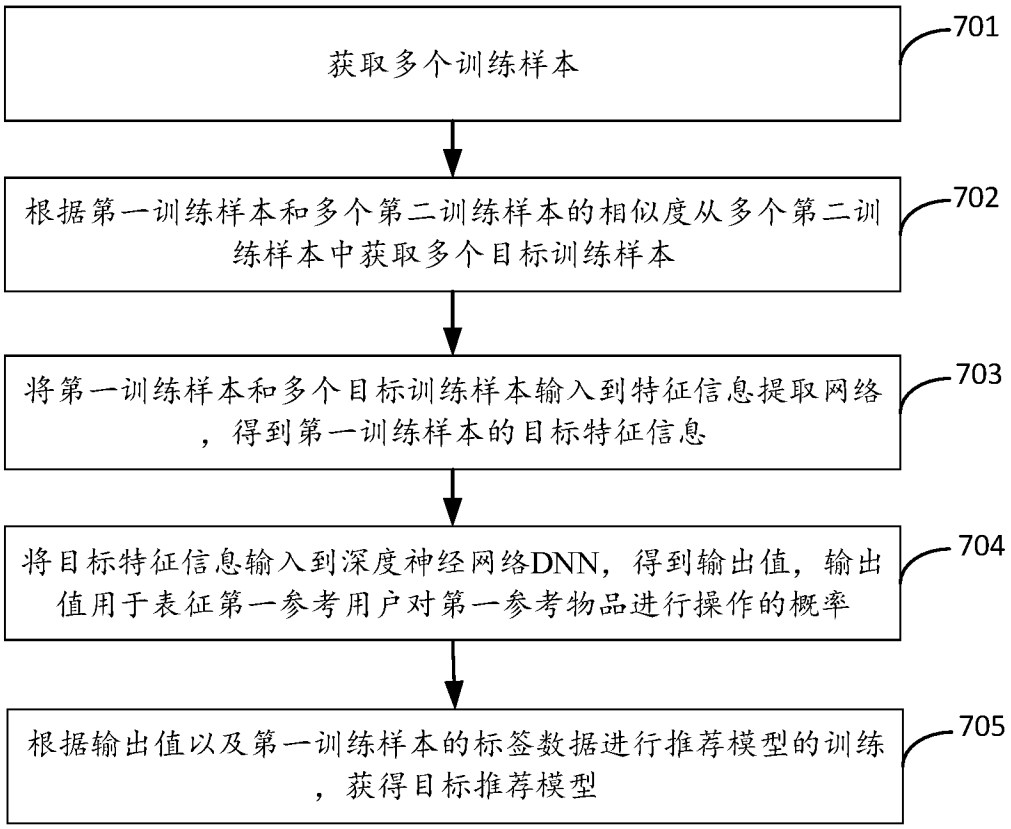

获取多个训练样本
701
根据第一训练样本和多个第二训练样本的相似度从多个第二训练样本中获取多个目标训练样本
702
将第一训练样本和多个目标训练样本输入到特征信息提取网络，得到第一训练样本的目标特征信息
703
将目标特征信息输入到深度神经网络DNN，得到输出值，输出值用于表征第一参考用户对第一参考物品进行操作的概率
704
根据输出值以及第一训练样本的标签数据进行推荐模型的训练，获得目标推荐模型
705

图 7

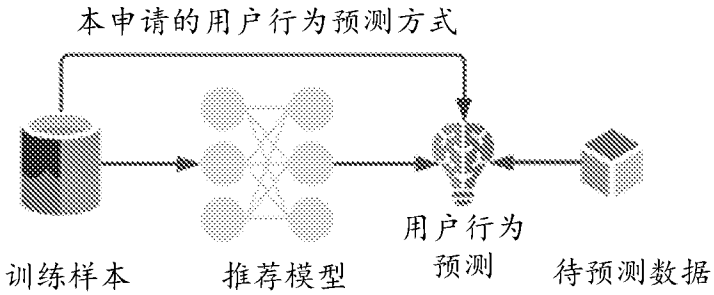

本申请的用户行为预测方式
训练样本
推荐模型
用户行为预测
待预测数据

图 8

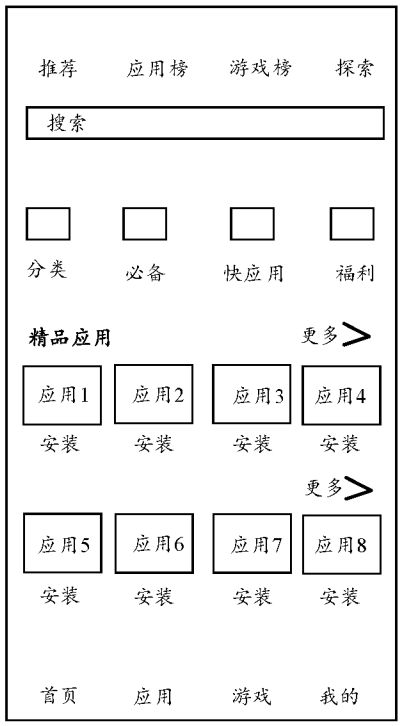

推荐
应用榜
游戏榜
探索
搜索
分类
必备
快应用
福利
精品应用
更多>
应用1
应用2
应用3
应用4
安装
安装
安装
安装
更多>
应用5
应用6
应用7
应用8
安装
安装
安装
安装
首页
应用
游戏
我的

图 9

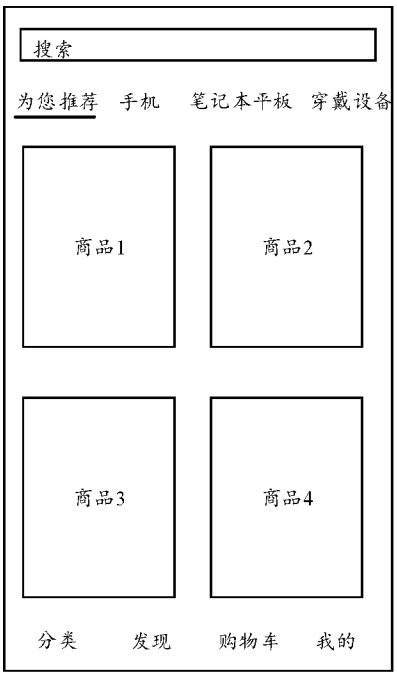

搜索
为您推荐
手机
笔记本平板
穿戴设备
商品1
商品2
商品3
商品4
分类
发现
购物车
我的

图 10

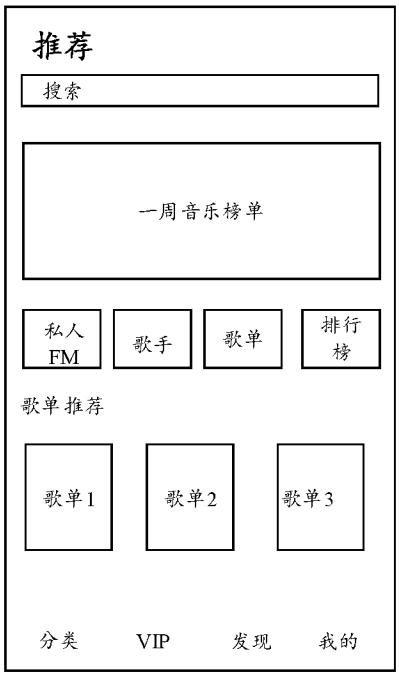

推荐
搜索
一周音乐榜单
私人FM
歌手
歌单
排行榜
歌单推荐
歌单1
歌单2
歌单3
分类
VIP
发现
我的

图 11

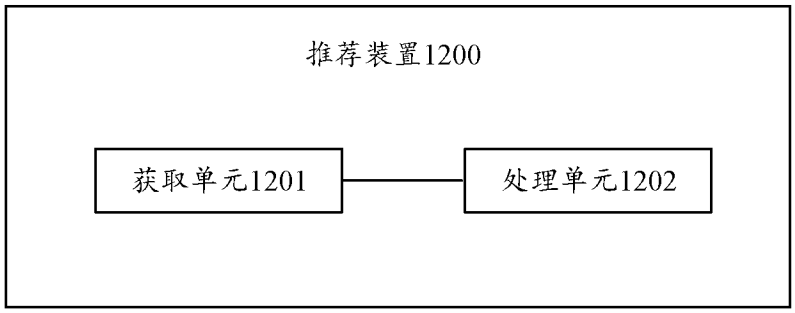

推荐装置1200
获取单元1201
处理单元1202

图 12

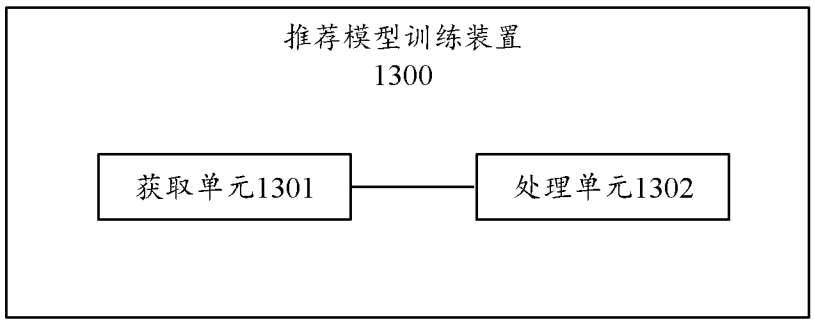

推荐模型训练装置
1300
获取单元1301
处理单元1302

图 13

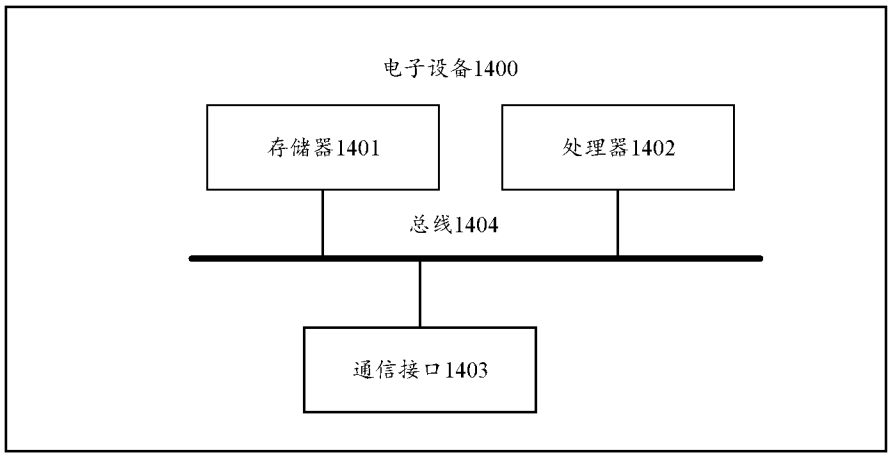


图 14

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2022/109297

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; G06Q; G06N

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)

CNKI, CNPAT, WPI, EPODOC: 推荐, 训练, 样本, 模型, 用户特征, 物品特征, 相似度, 深度神经网络, recommendation, training, sample, model, user information, object information, similarity, deep neural networks, DNN

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
PX	CN 113722583 A (HUAWEI TECHNOLOGIES CO., LTD.) 30 November 2021 (2021-11-30) claims 1-22, description, paragraphs [0002]-[0245], and figures 1-14	1-24
Y	CN 111259263 A (TENCENT CLOUD COMPUTING (BEIJING) CO., LTD.) 09 June 2020 (2020-06-09) description, paragraphs [0100]-[0251], and figures 1-6	1-24
Y	CN 109299396 A (NORTHEAST NORMAL UNIVERSITY) 01 February 2019 (2019-02-01) description, paragraphs [0002]-[0003] and [0056]-[0107], and figures 1-3	1-24
A	CN 111881363 A (BEIJING UNIVERSITY OF TECHNOLOGY) 03 November 2020 (2020-11-03) entire document	1-24
A	CN 112632403 A (BEIJING BAIDU NETCOM SCIENCE AND TECHNOLOGY CO., LTD.) 09 April 2021 (2021-04-09) entire document	1-24
A	WO 2021081962 A1 (HUAWEI TECHNOLOGIES CO., LTD.) 06 May 2021 (2021-05-06) entire document	1-24
A	US 2020272913 A1 (HUAWEI TECHNOLOGIES CO., LTD.) 27 August 2020 (2020-08-27) entire document	1-24

☑ 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
29 September 2022	**26 October 2022**

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

International application No.
PCT/CN2022/109297

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	薛峰 等 (XUE, Feng et al.). "基于深度神经网络和加权隐反馈的个性化推荐 (Personalized Recommendation Algorithm Based on Deep Neural Network and Weighted Implicit Feedback)" *模式识别与人工智能 (Pattern Recognition and Artificial Intelligence)*, 30 April 2020 (2020-04-30), entire document	1-24

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2022/109297

Patent document cited in search report	Publication date (day/month/year)	Patent family member(s)	Publication date (day/month/year)
CN 113722583 A	30 November 2021	None	
CN 111259263 A	09 June 2020	None	
CN 109299396 A	01 February 2019	None	
CN 111881363 A	03 November 2020	None	
CN 112632403 A	09 April 2021	None	
WO 2021081962 A1	06 May 2021	US 2021248651 A1	12 August 2021
		EP 3862893 A1	11 August 2021
		CN 113508378 A	15 October 2021
US 2020272913 A1	27 August 2020	CN 109902222 A	18 June 2019
		WO 2020107806 A1	04 June 2020

国际检索报告

国际申请号
PCT/CN2022/109297

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; G06Q; G06N

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））

CNKI,CNPAT,WPI,EPODOC:推荐，训练，样本，模型，用户特征，物品特征，相似度，深度神经网络，recommendation, training, sample, model, user information, object information, similarity, deep neural networks, DNN

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 113722583 A (华为技术有限公司) 2021年11月30日 (2021-11-30) 权利要求1-22，说明书第[0002]-[0245]段，图1-14	1-24
Y	CN 111259263 A (腾讯云计算北京有限责任公司) 2020年6月9日 (2020-06-09) 说明书第[0100]-[0251]段，图1-6	1-24
Y	CN 109299396 A (东北师范大学) 2019年2月1日 (2019-02-01) 说明书第[0002]-[0003]，[0056]-[0107]段，图1-3	1-24
A	CN 111881363 A (北京工业大学) 2020年11月3日 (2020-11-03) 全文	1-24
A	CN 112632403 A (北京百度网讯科技有限公司) 2021年4月9日 (2021-04-09) 全文	1-24
A	WO 2021081962 A1 (华为技术有限公司) 2021年5月6日 (2021-05-06) 全文	1-24
A	US 2020272913 A1 (HUAWEI TECHNOLOGIES CO., LTD.) 2020年8月27日 (2020-08-27) 全文	1-24

☑ 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9月29日	2022年10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 刘炯 电话号码 86-(10)-53961738

国际检索报告

国际申请号
PCT/CN2022/109297

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	薛峰 等. "基于深度神经网络和加权隐反馈的个性化推荐" 模式识别与人工智能, 2020年4月30日 (2020 - 04 - 30), 全文	1-24

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2022/109297

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	113722583	A	2021年11月30日	无			
CN	111259263	A	2020年6月9日	无			
CN	109299396	A	2019年2月1日	无			
CN	111881363	A	2020年11月3日	无			
CN	112632403	A	2021年4月9日	无			
WO	2021081962	A1	2021年5月6日	US	2021248651	A1	2021年8月12日
				EP	3862893	A1	2021年8月11日
				CN	113508378	A	2021年10月15日
US	2020272913	A1	2020年8月27日	CN	109902222	A	2019年6月18日
				WO	2020107806	A1	2020年6月4日