

业务处理方法、装置及计算机设备、存储介质、程序产品

相关申请

本申请要求 2023 年 10 月 20 日申请的，申请号为 2023113635030，名称为“业务处理方法、装置及计算机设备、存储介质、程序产品”的中国专利申请的优先权，在此将其全文引入作为参考。

技术领域

本申请涉及计算机技术领域，尤其涉及人工智能技术领域，具体涉及一种业务处理方法、装置及计算机设备、存储介质、程序产品。

背景技术

通常情况下，当业务用户进行业务访问时，业务服务方可以进行业务处理，业务处理具体可以是指向业务用户推荐业务中与业务用户关联度较高的物品，也就是说，业务服务方可以向访问业务的业务用户进行物品推荐。目前，业务服务方可以使用一些人工智能模型（例如，推荐模型）进行业务处理，然而，这些人工智能模型通常具有大量的模型参数，人工智能模型的推理速度慢，从而，导致业务处理效率不高。

发明内容

本申请实施例提供了一种业务处理方法、装置及计算机设备、存储介质、程序产品。一方面，本申请实施例提供了一种业务处理方法，由计算机设备执行，包括：

获取为业务用户召回的目标物品的提示信息，所述目标物品的提示信息包括所述业务用户的信息、与所述业务用户关联的至少一个关联物品的特征、以及所述目标物品的特征；

在模型应用推理框架下调用物品处理模型，基于所述目标物品的提示信息预测所述业务用户与所述目标物品之间的关联度；所述物品处理模型在模型训练推理框架下训练得到，并被复用至所述模型应用推理框架中；所述物品处理模型在所述模型应用推理框架中的处理速度高于所述物品处理模型在所述模型训练推理框架中的处理速度；及

基于所述业务用户与所述目标物品之间的关联度，对所述目标物品进行业务处理。

相应地，本申请实施例提供了一种业务处理装置，该业务处理装置包括：

获取单元，用于获取为业务用户召回的目标物品的提示信息，所述目标物品的提示信息包括所述业务用户的信息、与所述业务用户关联的至少一个关联物品的特征、以及所述目标物品的特征；

处理单元，用于在模型应用推理框架下调用物品处理模型，基于所述目标物品的提示信息预测所述业务用户与所述目标物品之间的关联度；所述物品处理模型在模型训练推理框架下训练得到，并被复用至所述模型应用推理框架中；所述物品处理模型在所述模型应用推理框架中的处理速度高于所述物品处理模型在所述模型训练推理框架中的处理速度；及

所述处理单元，还用于基于所述业务用户与所述目标物品之间的关联度，对所述目标物品进行业务处理。

相应地，本申请实施例提供一种计算机设备，该计算机设备包括：

处理器，适于实现计算机程序；

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，计算机程序适于由处理器加载并执行上述的业务处理方法。

相应地，本申请实施例提供一种计算机可读存储介质，该计算机可读存储介质存储有计算机程序，该计算机程序被计算机设备的处理器读取并执行时，使得计算机设备执行上述的业务处理方法。

相应地，本申请实施例提供了一种计算机程序产品，该计算机程序产品包括计算机程序，该计算机程序存储在计算机可读存储介质中。计算机设备的处理器从计算机可读存储介质读取该计算机程序，处理器执行该计算机程序，使得该计算机设备执行上述的业务处

理方法。

本申请的一个或多个实施例的细节在下面的附图和描述中提出。本申请的其它特征、目的和优点将从说明书、附图以及权利要求书变得明显。

附图说明

为了更清楚地说明本申请实施例或传统技术中的技术方案，下面将对实施例或传统技术描述中所需要使用的附图作简单地介绍，显而易见地，下面描述中的附图仅仅是本申请的实施例，对于本领域普通技术人员来讲，在不付出创造性劳动的前提下，还可以根据公开的附图获得其他的附图。

图 1 是本申请实施例提供的一种业务处理系统的架构示意图；

图 2 是本申请实施例提供的一种物品处理模型的训练过程和测试过程示意图；

图 3 是本申请实施例提供的一种业务处理方法的流程示意图；

图 4 是本申请实施例提供的一种物品处理模型的训练过程示意图；

图 5 是本申请实施例提供的一种模型应用推理框架下改造前的物品处理模型的结构示意图；

图 6 是本申请实施例提供的一种模型应用推理框架下改造后的物品处理模型的结构示意图；

图 7 是本申请实施例提供的另一种物品处理方法的流程示意图；

图 8 是本申请实施例提供的一种并发机制的示意图；

图 9 是本申请实施例提供的另一种并发机制的示意图；

图 10 是本申请实施例提供的再一种并发机制的示意图；

图 11 是本申请实施例提供的一种业务处理装置的结构示意图；

图 12 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

下面将结合本申请实施例中的附图，对本申请实施例中的技术方案进行清楚、完整地描述，显然，所描述的实施例仅仅是本申请一部分实施例，而不是全部的实施例。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。

本申请实施例涉及业务，业务具体可以是指物品推荐业务，物品推荐业务的业务处理逻辑为在业务用户进行业务访问时，在为业务用户召回的多个候选物品中，向业务用户推荐与业务用户之间关联度较高的候选物品。本申请实施例提及的物品可以包括但不限于以下任一种：电子商品（例如，智慧零售商品以及电商商品）以及媒体信息（例如，视频、音频、图像、文章以及网址），等等。

业务服务方可以使用人工智能模型进行业务处理，具体来说，业务服务方可以使用人工智能模型对业务用户与为业务用户召回的候选物品进行关联预测，得到业务用户与候选物品之间的关联度，从而，可以基于业务用户与候选物品之间的关联度，向业务用户推荐与业务用户之间关联度较高的候选物品。

目前，在人工智能领域中，随着 ChatGPT（Chat Generative Pre-trained Transformer，聊天生成预训练转换器）大火，越来越多的 LLM（Large Language Model，大规模生成式语言模型）涌现。本申请实施例不对大规模生成式语言模型的类型进行限定，大规模生成式语言模型是一种模型框架为 Encoder-Decoder（编码-解码）框架、Encoder-only 框架或 Decoder-only 框架的生成式语言模型；其中，Encoder-Decoder 框架是深度学习中非常常见的一个模型框架，一个最显著的特征就是它是一个 End-to-End（端到端）学习的算法，这样的模型也被叫做 Sequence to Sequence Learning（序列到序列的学习），所谓编码是指将输入序列转化成一个固定长度的向量，所谓解码是指将之前生成的固定向量再转化成输出序列；Encoder-only 框架是指模型具备编码的功能；Decoder-only 框架是指模型具备解码的功能。本申请实施例以大规模生成式语言模型是模型框架为 Encoder-Decoder 框架的生成

式语言模型为例进行说明，例如，大规模生成式语言模型可以是模型框架为 Encoder-Decoder 框架的 MT5 模型。MT5 模型是一种多语言自然语言处理模型，该模型采用了 Transformer 模型架构，并在大规模多语言语料库上进行了预训练。Transformer 模型是一种自然语言处理模型，Transformer 模型可以根据输入序列中各个单词（token）之间的关系，学习各个单词的上下文语义，Transformer 模型基于 Encoder-Decoder 框架和 Attention（注意力）机制就能达到很好的语义学习效果，Transformer 模型最大的优点是可以高效地并行化。单词（token）是输入序列的组成元素，可以将句子、段落以及文章等类型的输入序列，分解为以单词（token）为单位的数据结构输入大规模生成式语言模型中。

如何将大规模生成式语言模型作为物品处理模型应用于物品推荐业务中，成为当前的研究热点。然而，大规模生成式语言模型目前无法满足物品推荐业务的物品推荐需求，具体来说：①大规模生成式语言模型具备大量的模型参数，模型推理速度慢，模型的推理时延高，模型推理性能限制了大规模生成式语言模型在实时的物品推荐业务中落地，不能满足物品推荐业务的实时推荐和并行推理需求；②大规模生成式语言模型生成的 token（单词）不能满足物品推荐业务的推荐打分排序需求；基于①和②可以看出，大规模生成式语言模型目前无法作为物品处理模型应用于物品推荐业务中。

针对大规模生成式语言模型推理性能差（即大规模生成式语言模型不能满足物品推荐业务的实时推荐需求）的问题，业界提出了以下三种解决方案：第一种解决方案，采用模型压缩或者模型蒸馏的方式，将大规模生成式语言模型的模型参数缩减，以牺牲模型精度的方式提升模型推理速度；第二种解决方案，对预训练模型进行微调，根据设定的阈值提前退出推理，减少模型网络层数；第三种解决方案，缓存大规模生成式语言模型中解码模块的中间结果，提升模型推理速度。但是在物品推荐业务的实践中发现，上述三种解决方案并不可取，对于第一种解决方案和第二种解决方案，以牺牲结果精度来提升推理速度，这对于应用到实时物品推荐中并不可取，因为牺牲结果精度后无法保障大规模生成式语言模型在实时物品推荐场景中的效果，会使得预估的 AUC 值下降较多；其中，预估的 AUC 是一个模型评价指标，能够用于二分类模型的评价，预估的 AUC 值越高，模型分类的准确性越高，预估的 AUC 值越低，模型分类的准确性越低。对于第三种解决方案，需要增加内存缓存，在资源有限的情况下，这种解决方案也无法提升大规模生成式语言模型的模型推理速度。

基于此，本申请实施例提出一种业务处理方法，该业务处理方法针对大规模生成式语言模型推理性能差（即大规模生成式语言模型不能满足物品推荐业务的实时推荐需求）的问题，提出一种大规模生成式语言模型的训练框架与推理框架分离的解决方案，大规模生成式语言模型在推理框架下的处理速度高于大规模生成式语言模型在训练框架下的处理速度，在大规模生成式语言模型的模型推理阶段使用能够使得模型处理速度提升的推理加速框架，且没有牺牲大规模生成式语言模型的模型精度，这样可以在保证模型效果的前提下，提升大规模生成式语言模型的模型推理速度，从而可以提升业务处理效率。

该业务处理方法针对大规模生成式语言模型不能满足物品推荐业务的推荐打分排序需求的问题，提出一种对大规模生成式语言模型的解码器进行优化改造的解决方案，优化改造后的大规模生成式语言模型能够对候选物品进行打分，也就是说，优化改造后的大规模生成式语言模型能够输出业务用户与候选物品之间的关联度，将业务用户与候选物品之间的关联度作为候选物品的打分，对候选物品进行排序推荐，使得大规模生成式语言模型能够适应物品推荐业务，成功作为物品处理模型应用于物品推荐业务中。

本申请实施例不对大规模生成式语言模型（即物品处理模型）使用的训练框架的类型进行限定，例如，训练框架可以是 pytorch 框架。类似地，本申请实施例不对大规模生成式语言模型（即物品处理模型）使用的推理框架的类型进行限定，推理框架可以是能够提升大规模生成式语言模型（即物品处理模型）的处理速度的任意推理加速框架；例如，推理框架可以是 Faster-Transformer 的推理加速框架，Faster-Transformer 是针对 Transformer 模

型推理提出的性能优化方案，Faster-Transformer 采用 c++和 cuda 编写，依赖高度优化的 cuBLAS；其中，c++（c plus plus）是一种计算机高级程序设计语言，cuda（Compute Unified Device Architecture，统一计算设备架构）是一种通用并行计算架构，cuBLAS（CUDA Basic Linear Algebra Subroutine Library）是 cuda 的基本线性代数子程序库。也就是说，本申请实施例可以在 pytorch 框架下适配物品推荐业务对大规模生成式语言模型（即物品处理模型）进行微调，模型推理阶段为了提升模型推理速度，可以采用基于 Faster-Transformer 的推理加速框架，复用微调后的大规模生成式语言模型（即物品处理模型），并在推理加速框架下对大规模生成式语言模型（即物品处理模型）的解码器优化输出打分。

此外，本申请实施例提出的业务处理方法还可以与云技术相结合，例如，云技术中的云计算可以通过分布式的计算模式，进一步提升业务处理方法中大规模生成式语言模型（即物品处理模型）的模型推理速度。云技术（Cloud Technology）是指在广域网或局域网内将硬件、软件以及网络等系列资源统一起来，实现数据的计算、储存、处理和共享的一种托管技术；云技术基于云计算商业模式应用的网络技术、信息技术、整合技术、管理平台技术、应用技术等的总称，可以组成资源池，按需所用，灵活便利。云计算（Cloud Computing）是一种计算模式，它将计算任务分布在大量计算机构成的资源池上，使各种应用系统能够根据需要获取计算力、存储空间和信息服务。提供资源的网络被称为“云”，“云”中的资源在使用者看来是可以无限扩展的，并且可以随时获取，按需使用，随时扩展，按使用付费。

下面结合附图介绍本申请实施例提供的业务处理系统，业务处理系统可以适于实现本申请实施例提供的业务处理方法。

如图 1 所示，业务处理系统可以包括推荐请求设备 101 和物品推荐设备 102，本申请实施例不对推荐请求设备 101 和物品推荐设备 102 之间的连接方式进行限定，推荐请求设备 101 和物品推荐设备 102 之间可以通过有线通信的方式建立直接的通信连接，或者，推荐请求设备 101 和物品推荐设备 102 之间可以通过无线通信的方式建立间接的通信连接。其中：

（1）推荐请求设备 101：

推荐请求设备 101 可以是业务用户使用的设备，推荐请求设备 101 中可以运行有业务对应的业务产品，业务用户可以通过推荐请求设备 101 中运行的业务产品访问业务；访问业务可以理解为是业务用户登录业务、业务用户浏览业务内容或业务用户执行业务操作，等等。本申请实施例不对业务产品的类型进行限定，业务产品可以是业务应用程序、业务小程序、业务浏览器以及业务网站，等等。推荐请求设备 101 可以用于向物品推荐设备 102 发起物品推荐请求，物品推荐请求可以用于请求向业务用户推荐与业务用户关联度较高的物品。物品推荐请求可以是业务用户访问业务时生成的；例如，当业务用户登录业务时可以生成物品推荐请求；又如，当业务用户浏览业务内容时可以生成物品推荐请求；再如，当业务用户执行业务操作（例如，物品推荐操作以及访问业务中物品的操作，等等）时可以生成物品推荐请求。推荐请求设备 101 还可以向业务用户展示推荐物品的相关信息（例如，物品名称、物品标题、物品图片、物品视频以及物品价格，等等）。

（2）物品推荐设备 102：

物品推荐设备 102 可以用于向业务用户推荐与业务用户关联度较高的物品。具体来说：S101，物品推荐设备 102 可以响应于推荐请求设备 101 发送的物品推荐请求，为业务用户召回多个候选物品；S102，针对多个候选物品中任一候选物品（可以表示为目标物品），物品推荐设备 102 可以在模型应用推理框架（模型应用推理框架是指前文提及的推理框架）下调用物品处理模型预测业务用户与候选物品之间的关联度；S103，在得到业务用户与每个候选物品之间的关联度之后，物品推荐设备 102 可以根据业务用户与每个候选物品之间的关联度，在多个候选物品中选择与业务用户之间关联度较高的物品，将与业务用户之间

关联度满足推荐条件的物品作为向业务用户推荐的推荐物品；S104，向推荐请求设备 101 返回推荐物品的相关信息。其中，关联度满足推荐条件，可以是关联度大于等于预设阈值，或者按照关联度降序排序后排序序数小于或等于预设排序序数。

对于物品推荐设备 102，在物品处理模型正式上线进行物品推荐之前，物品处理模型需要经历模型训练阶段和模型测试阶段，在物品处理模型通过模型测试后，可以正式上线进行物品推荐。

如图 2 所示，在模型训练阶段中：S201a，可以准备用于对物品处理模型进行训练的训练样本；S202a，为了适应物品推荐业务，可以对初始的物品处理模型进行改造，使得初始的物品处理模型具备对物品进行打分的能力；S203a，可以在模型训练推理框架（模型训练推理框架是指前文提及的训练框架）下基于训练样本对初始的物品处理模型进行训练，得到物品处理模型。

如图 2 所示，在模型测试阶段：S201b，可以准备用于对物品处理模型进行测试的测试样本；S202b，可以将物品处理模型与模型应用推理框架进行部署适配，部署适配是指为了适应物品推荐业务，与模型训练阶段类似，对模型应用推理框架下的物品处理模型进行改造，使得模型应用推理框架下的物品处理模型具备对物品进行打分的能力；S203b，对模型应用推理框架下的物品处理模型进行线上部署，具体来说，在将物品处理模型与模型应用推理框架进行部署适配后，可以编译框架 c++ 代码，获取 lib 文件（一种静态数据连接库），线上部署是指根据编译产出的 lib 文件，结合模型推理代码，部署线上模型推理服务，部署中需要注意 cuda 版本、gcc（一种编译器）等编译环境版本一致；S204b，对物品处理模型进行模型校准，由于本申请实施例中模型训练阶段和模型推理阶段采用的框架不同，因此需要进行模型校准，模型校准是指基于同一批测试样本，在模型训练阶段采用的模型训练推理框架（例如，pytorch 框架）和模型推理阶段采用的模型应用推理框架（例如，Faster-Transformer 框架）下对物品处理模型进行评估，若两个框架下物品处理模型的评估结果（评估结果具体可以是指预估的 AUC 值）之间的差值小于阈值 thres（阈值 thres 可以根据经验值进行设定，例如，阈值 thres 通常可以为 0.00001），则可以确定对物品处理模型测试通过；S205b，在确定对物品处理模型测试通过后，可以对物品处理模型进行上线，接入线上真实流量，在模型应用推理框架下调用物品处理模型进行物品推荐。

需要说明的是，推荐请求设备 101 可以是终端，物品推荐设备 102 可以是终端或服务器；本申请实施例提及的终端可以包括但不限于以下任一种：智能手机、平板电脑、笔记本电脑、台式计算机、智能手表、智能家电、智能语音交互设备、车载终端以及飞行器，但并不局限于此；本申请实施例提及的服务器可以是单独的物理服务器，也可以是多个物理服务器构成的服务器集群或者分布式系统，还可以是提供云服务、云数据库、云计算、云函数、云存储、网络服务、云通信、中间件服务、域名服务、安全服务、CDN（Content Delivery Network，内容分发网络）、以及大数据和人工智能平台等基础云计算服务的云服务器，本申请实施例对此不进行限定。此外，图 1 所示的业务处理系统以推荐请求设备 101 和物品推荐设备 102 集成于不同的设备为例进行示意，推荐请求设备 101 和物品推荐设备 102 也可以集成于同一设备中，例如，推荐请求设备 101 和物品推荐设备 102 可以集成于同一终端中，在同一终端中，推荐请求设备 101 可以理解为是终端的前端，物品推荐设备 102 可以理解为是终端的后台服务。

可以理解的是，图 1 所示实施例的业务处理系统是为了更加清楚的说明本申请实施例的技术方案，并不构成对于本申请实施例提供的技术方案的限定，本领域普通技术人员可知，随着系统架构的演变和新业务场景的出现，本申请实施例提供的技术方案对于类似的技术问题，同样适用。另外，本申请实施例中相关数据收集处理（例如，收集业务用户的信息）在实例应用时应该严格根据相关法律法规的要求，获取个人信息主体的知情同意或单独同意，并在法律法规及个人信息主体的授权范围内，开展后续数据使用及处理行为。

基于上述对业务处理方法和业务处理系统的介绍，下面对本申请实施例提供的业务处

理方法的具体应用进行介绍。

从业务中的物品类型上划分，本申请实施例提供的业务处理方法可以应用于智慧零售、电商推荐、信息流推荐等业务中。具体来说，在智慧零售业务中，向业务用户推荐的物品例如可以是智慧零售商品，可以分析业务用户与为业务用户召回的每个智慧零售商品之间的关联度，基于业务用户与每个智慧零售商品之间的关系，向业务用户推荐与业务用户关联度较高的智慧零售商品。在电商推荐场景中，向业务用户推荐的物品例如可以是电商商品，可以分析业务用户与为业务用户召回的每个电商商品之间的关联度，基于业务用户与每个电商商品之间的关系，向业务用户推荐与业务用户关联度较高的电商商品。在信息流推荐业务中，向业务用户推荐的物品例如可以是视频、音频、图像以及文章等媒体信息，可以分析业务用户与为业务用户召回的每个媒体信息之间的关联度，基于业务用户与每个媒体信息之间的关系，向业务用户推荐与业务用户关联度较高的媒体信息。

从业务复杂程度上划分，本申请实施例提供的业务处理方法可以应用于数据较为稀疏（例如，业务中的物品数量较少、业务中物品的相关信息较少，等等）、以及传统排序推荐模型推荐效果较差的简单业务中。本申请实施例提供的业务处理方法也可以应用于涉及多个业务场景或多个接口的复杂业务中。多个业务场景可以理解为业务中同时存在多个业务场景，可以在多个业务场景中向业务用户进行物品推荐，例如，业务可以包含电商推荐业务场景和信息流推荐业务场景，可以在电商推荐业务场景向业务用户进行电商商品推荐，在信息流推荐业务场景中向业务用户进行信息流推荐，并且多个业务场景可以并行进行物品推荐。多个接口可以理解为业务中通过不同的接口接入了多个子业务，可以在每个子业务中向业务用户进行物品推荐，各个子业务的子业务服务可以并行进行物品推荐；或者，多个接口可以理解为业务中通过不同的接口接入了多个业务用户，可以为每个业务用户进行物品推荐，各个业务用户可以并行进行物品推荐，例如，业务中可以同时接入至少 5 个业务用户。在模型效果持平的基础上，采用本申请实施例的推理加速框架可以比原生框架推理耗时减少 5 倍。

可以看出，本申请实施例可以利用自然语言模型中丰富的语言知识，针对物品推荐业务进行微调，提高大语言模型对该物品推荐场景的学习能力，并且本申请实施例支持高并发推理，可以接入多个业务场景、多个子业务或多个业务用户，能够解决性能瓶颈，推动大语言模型在工业界的落地。

下面结合附图，对本申请实施例提供的业务处理方法进行详细介绍。

本申请实施例提供一种业务处理方法，该业务处理方法介绍物品处理模型在模型训练阶段和模型推理阶段适应业务（即物品推荐业务）进行的改造。该业务处理方法可以由计算机设备执行，计算机设备例如可以是图 1 所示的业务处理系统中的物品推荐设备 102，或者，计算机设备例如可以是推荐请求设备 101 和物品推荐设备 102 的集成设备。如图 3 所示，该业务处理方法可以包括但不限于以下步骤 S301-步骤 S303：

S301，获取为业务用户召回的目标物品的提示信息，目标物品的提示信息包括业务用户的信息、与业务用户关联的至少一个关联物品的特征、以及目标物品的特征。

在业务方提供的业务（即物品推荐业务）中，当业务用户访问业务时，业务服务方可以为业务用户召回多个候选物品，需要通过物品处理模型分析业务用户与每个候选物品之间的关联度，然后，可以基于业务用户与每个候选物品之间的关联度，在多个候选物品中选择向业务用户推荐的推荐物品，以及根据推荐物品对业务用户进行物品推荐。为了便于理解业务用户与每个候选物品之间关联度的分析过程，在此以多个候选物品中的任一个候选物品（任一个候选物品可以表示为目标物品）为例，介绍业务用户与目标物品之间关联度的分析过程。

具体来说，业务用户与目标物品之间的关联度是基于目标物品的提示信息（prompt）进行分析的，目标物品的提示信息作为物品处理模型的输入信息，可以获取目标物品的提

示信息，目标物品的提示信息可以包括业务用户的信息、与业务用户关联的至少一个关联物品的特征、以及目标物品的特征；其中，与业务用户关联的至少一个关联物品是指业务用户在业务中历史访问的历史物品，物品（包括关联物品、目标物品或候选物品）的特征是指用于表征物品特点的信息。

不同于传统推荐模型的输入信息，物品处理模型属于生成式语言模型，物品处理模型的输入信息（即提示信息）是由自然语言构成，目标物品的提示信息可以用于询问业务用户在历史访问关联物品的情况下，业务用户是否会访问目标物品。物品处理模型的分析原理为，通过理解目标物品的提示信息的语义，分析目标物品的特征与至少一个关联物品的特征之间的相似关系，分析目标物品的特征与业务用户的信息之间的关联关系，生成业务用户与目标物品之间的预测关联标签（answer）和预测关联标签下的关联度，预测关联标签可以用于表征业务用户访问目标物品的意愿（例如，意愿为是（即业务用户会访问目标物品）或意愿为否（即业务用户不会访问目标物品）），关联度可以用于表征业务用户在相应的预测关联标签下访问目标物品的概率，可以作为对目标物品进行排序的依据。

为业务用户召回的各个候选物品的提示信息的构造方式是相同的，在此介绍目标物品的提示信息的构造方式。首先，可以从数据中心获取业务用户历史访问的关联物品的物品标识序列（item1, item2, ..., itemn）、目标物品的物品标识（now_click_item）以及业务用户的信息；其中，物品标识可以用于唯一标识其对应的物品，业务用户在业务中对关联物品的访问操作至少可以包括点击浏览操作和购买操作，从而，业务用户历史访问的关联物品的物品标识序列可以包括业务用户历史点击的关联物品的点击标识序列（click_item1, click_item2, ..., click_itemn），以及业务用户历史购买的关联物品的购买标识序列（buy_item1, buy_item2, ..., buy_itemn）；业务用户的信息至少可以包括业务用户的标识和业务用户的特征，业务用户的标识可以用于唯一标识业务用户，业务用户的特征是指用于表征业务用户特点的信息。然后，可以根据业务用户历史访问的关联物品的物品标识序列（item1, item2, ..., itemn），从物料库中获取业务用户历史访问的关联物品的特征，以及，可以根据目标物品的物品标识（now_click_item），从物料库中获取目标物品的特征；其中，以物品的特征为物品标题为例，业务用户历史访问的关联物品的特征可以包括业务用户历史点击的关联物品的点击标题序列（title1, title2, title3, ..., titlen），以及业务用户历史购买的关联物品的购买标题序列（buy_title1, buy_title2, ..., buy_titlen）；以物品的特征为物品标题为例，目标物品的特征为目标物品的标题（now_click_title）。基于此，构造出的目标物品的提示信息可以包括以下任一种：

（1）基础的提示信息：基础的提示信息可以包括业务用户的标识、业务用户历史访问的关联物品的特征、以及目标物品的特征；基础的提示信息可以用于询问业务用户在历史访问关联物品的情况下，业务用户是否会访问目标物品。例如，基础的提示信息可以表示为：业务用户点击过关联物品：[title1],[title2],[title3],...,[titlen]，则业务用户是否会点击目标物品[now_click_title]。

（2）包含时效特征的提示信息：包含时效特征的提示信息可以包括业务用户的标识、业务用户在不同历史时段访问的关联物品的特征、以及目标物品的特征；包含时效特征的提示信息可以用于询问业务用户在不同历史时段访问相应关联物品的情况下，业务用户是否会访问目标物品。例如，包含时效特征的提示信息可以表示为：业务用户点击过关联物品：[title1],[title2],[title3],...,[titlen]，并且最近点击过关联物品：[title1],[title2],...，那么业务用户是否会点击目标物品[now_click_title]。

（3）包含时效特征+业务用户的特征的提示信息：包含时效特征+业务用户的特征的提示信息可以包括业务用户的特征、业务用户在不同历史时段访问的关联物品的特征、以及目标物品的特征；包含时效特征+业务用户的特征的提示信息，可以用于询问具备特征的业务用户在不同历史时段访问相应关联物品的情况下，业务用户是否会访问目标物品。例如，包含时效特征+业务用户的特征的提示信息可以表示为：具备[特征 1],[特征 2],...的业

务用户，曾经点击过关联物品：[title1],[title2],[title3],...,[titlen]，并且最近点击过关联物品 [title1],[title2]...，那么业务用户是否会点击目标物品[now_click_title]。

(4) 包含时效特征+业务用户的特征+访问特征的提示信息：包含时效特征+业务用户的特征+访问特征的提示信息可以包括业务用户的特征、业务用户在不同历史时段和不同访问操作下访问的关联物品的特征、以及目标物品的特征；包含时效特征+业务用户的特征+访问特征的提示信息可以用于询问具备业务用户的特征的业务用户在不同历史时段和不同访问操作下访问相应关联物品的情况下，业务用户是否会访问目标物品。例如，包含时效特征+业务用户的特征+访问特征的提示信息可以表示为：具备[业务用户的特征 1],[业务用户的特征 2],...的业务用户，曾经点击过关联物品：[title1],[title2],[title3],...,[titlen]，并且最近购买过关联物品 [title1],[title2]...，那么业务用户是否会点击目标物品[now_click_title]。

需要说明的是，在实际的物品推荐过程中，候选物品可以采用以上四种形式中任一种形式的提示信息；或者，在实际的物品推荐过程中，候选物品可以采用目标形式的提示信息，目标形式是在模型校准环节中，从以上四种形式中选择的能够使得模型具有较好推理效果（例如，具有较高推理准确度）的形式，这样有利于提升物品推荐的准确率，提升业务处理准确率。

S302，在模型应用推理框架下调用物品处理模型，基于目标物品的提示信息预测业务用户与目标物品之间的关联度；物品处理模型在模型训练推理框架下训练得到，并被复用至模型应用推理框架中；物品处理模型在模型应用推理框架中的处理速度高于物品处理模型在模型训练推理框架中的处理速度。

其中，模型应用推理框架，是物品处理模型在应用阶段下运行所使用的推理框架，推理框架是一种用于对已训练完成的算法模型进行预测的技术框架。模型训练推理框架，是物品处理模型在训练阶段下训练所使用的推理框架。业务用户与目标物品之间的关联度表征业务用户与目标物品之间关联程度的大小。

在获取到目标物品的提示信息后，可以在模型应用推理框架下调用物品处理模型，基于目标物品的提示信息，对业务用户与目标物品之间进行关联预测，得到业务用户与目标物品之间的关联度。其中，物品处理模型在模型训练推理框架下训练得到，并被复用至模型应用推理框架中，也就是说，物品处理模型在训练过程中采用的训练框架为模型训练推理框架，物品处理模型在推理过程中采用的推理框架为模型应用推理框架；物品处理模型在模型应用推理框架中的处理速度高于物品处理模型在模型训练推理框架中的处理速度。

物品处理模型是基于 Transformer Encoder-Decoder 架构的大规模生成式语言模型，而在物品推荐业务所需的模型是分类模型，并且物品推荐业务需要物品处理模型输出业务用户是否访问候选物品的概率（即需要输出业务用户与候选物品之间的关联度），因此，本申请实施例对物品处理模型进行改造，将原有的大规模生成式语言模型改造为适配物品推荐业务的分类模型。因为本申请实施例在物品处理模型的训练过程中采用的训练框架和在物品处理模型的推理过程中采用的推理框架不同，所以，本申请实施例在训练框架和推理框架中分别对物品处理模型进行适应性改造，下面分别介绍物品处理模型在训练过程中的适应性改造，以及物品处理模型在推理过程中的适应性改造：

(1) 物品处理模型的训练过程：

物品处理模型可以是在大规模多语言语料库上进行预训练得到的模型，本申请实施例提及的对物品处理模型进行训练，可以理解为是对物品处理模型进行微调。如图 4 所示，在物品处理模型的训练过程中，物品处理模型可以包括编码器（Encoder）和解码器（Decoder），编码器（Encoder）可以用于对物品处理模型的输入信息进行编码处理，解码器（Decoder）可以用于对编码器的编码结果进行解码处理，生成预测关联标签。物品处理模型在训练过程中的适应性改造可以包括：本申请实施例对物品处理模型的解码器（Decoder）部分进行改造，将原有的生成式语言模型改造为适配物品推荐业务的分类模型，原有的解码器（Decoder）输出为当前轮次的 token（即预测关联标签），本申请实施例对解

码器 (Decoder) 进行改造, 增加概率预测模块 (softmax), 通过概率预测模块 (softmax) 输出预估 token 的概率值 (即预测关联标签下的关联度)。

经过适应性改造, 物品处理模型的训练过程可以包括: 获取初始处理模型 (即前文所提及的初始的物品处理模型) 和样本数据, 样本数据中可以包括用于对初始处理模型进行训练的训练集, 训练集可以包括多个训练样本, 每个训练样本可以包括训练物品的提示信息 (prompt) 和训练物品对应的标记关联标签 (原始样本 label 或 answer), 可以在模型训练推理框架下基于训练集对初始处理模型进行训练, 得到物品处理模型。

更为详细地, 物品处理模型的训练过程是迭代进行的, 迭代进行可以理解为, 采用第一个训练样本对初始处理模型进行第一次训练, 采用第二个训练样本对第一次训练得到的物品处理模型进行第二次训练, 以此类推, 直至达到训练终止条件, 将达到训练终止条件的物品处理模型作为最终训练得到的物品处理模型。迭代训练过程中的任一次训练过程可以包括: 将训练样本中训练物品的提示信息 (可以称为样本 prompt) 输入物品处理模型, 训练物品的提示信息经分词模块 (tokenizer) 划分为多个单词 (如图 4 所示的单词 1-单词 n); 经过物品处理模型的编码器 (Encoder) 输出训练物品的提示信息中各个单词的上下文语义信息 (embedding), 然后经过解码器 (Decoder) 生成预估 token (即预测关联标签) 和 token 概率 (即预测关联标签下的关联度); 然后, 可以根据预估 token (即预测关联标签)、训练物品的标记关联标签以及 token 概率 (即预测关联标签下的关联度) 计算损失信息 (loss), 本申请实施例不对计算损失信息所采用的损失函数类型进行限定, 例如, 计算损失信息所采用的损失函数可以为交叉熵损失函数; 然后, 可以根据损失信息 (loss) 经过模型反向传播, 调整物品处理模型中编码器 (Encoder) 和解码器 (Decoder) 的模型参数。

(2) 物品处理模型的推理过程:

物品处理模型在训练过程和推理过程中采用不同的框架, 在训练得到物品处理模型后, 物品处理模型被部署于推理框架 (即模型应用推理框架) 中, 与物品处理模型的训练过程类似, 在物品处理模型的推理过程中, 需要将原有的大规模生成式语言模型改造为适配物品推荐业务的分类模型。如图 5 所示, 在物品处理模型的推理过程中, 在模型应用推理框架下改造前的物品处理模型的结构, 可以包括:

①n1 个编码器 (Encoder): n1 个编码器 (Encoder) 的结构相同, 编码器 (Encoder) 可以用于对编码输入信息 (encoder inputs) 进行编码处理。进一步地, 任一个解码器可以包括自注意机制模块 (Self-Attention) 和前馈神经网络模块 (Feed Forward Network); 其中, 自注意机制模块 (Self-Attention) 可以用于学习编码输入信息中的各个单词之间的关联性, 前馈神经网络模块可以用于对自注意机制模块的输出结果进行非线性映射, 对自注意机制模块的输出结果进行进一步学习。

②解码器 (Decoding): 解码器可以包括 n2 个解码模块 (decoder) 和标签预测模块 (beam search), n2 个解码模块 (decoder) 的结构相同; 解码模块可以用于结合解码输入信息 (decoding inputs) 和 n1 个编码器的编码结果 (encoder outputs) 进行解码处理; 标签预测模块 (beam search) 可以用于基于 n2 个解码模块 (decoder) 的解码结果进行关联标签预测, 生成预测关联标签。进一步地, 任一个解码模块 (decoder) 中可以包括自注意机制模块 (Self-Attention)、交叉注意力机制模块 (encoder-decoder Cross-Attention) 以及前馈神经网络模块 (Feed Forward Network); 其中, 自注意机制模块 (Self-Attention) 可以用于学习解码输入信息中各个单词之间的关联性, 交叉注意力机制模块 (encoder-decoder Cross-Attention) 可以学习编码输入信息与解码输入信息之间的关联性, 前馈神经网络模块可以用于对交叉注意力机制模块的输出结果进行非线性映射, 对交叉注意力机制模块的输出结果进行进一步学习。

如图 6 所示, 物品处理模型在推理过程中的适应性改造采用虚线框进行标示, 物品处理模型在推理过程中的适应性改造可以包括: 对于物品处理模型, 在第一架构下物品处理模型解码器 (Decoding) 只输出生成的 token (即预测关联标签), 并不支持分类模型及输

出 token 概率（即预测关联标签下的关联度）等功能。为了与训练过程中物品处理模型的改造一致，可以采用概率预测模块（softmax）对标签预测模块（beam search）进行改造，改造后概率预测模块（softmax）和标签预测模块（beam search）合并为预测模块，可以获取原始 token 的 softmax 值作为 token 概率并输出，即输出为[output id, output prob]，其中，output id 为预测关联标签，output prob 为预测关联标签下的关联度。

经过适应性改造，物品处理模型的推理过程可以包括：以目标物品为例，物品处理模型的编码输入信息为目标物品的提示信息，目标物品的提示信息输入物品处理模型后被划分为多个单词，可以获取目标物品的提示信息中各个单词的表征信息，任一单词的表征信息是指可以用于唯一表征单词的向量表示；在模型应用推理框架下调用编码器基于各个单词的表征信息之间的关联性，对每个单词进行语义分析，得到每个单词的上下文语义信息；在模型应用推理框架下调用解码模块（具体可以是解码模块的交叉注意力机制模块和前馈神经网络模块）基于每个单词的上下文语义信息与解码输入信息的上下文语义信息（解码输入信息的上下文语义信息可以通过解码模块的自注意力机制模块获得的）之间的关联性，生成解码输入信息的解码表示；可以在模型应用推理框架下调用预测模块基于解码输入信息的解码表示，对业务用户与目标物品进行关联预测，得到业务用户与目标物品之间的关联度。

更为详细地，在模型应用推理框架下调用预测模块基于解码输入信息的解码表示，对业务用户与目标物品进行关联预测，得到业务用户与目标物品之间的关联度的过程，可以包括：在模型应用推理框架下调用预测模块中的标签预测模块基于解码输入信息的解码表示，对业务用户与目标物品进行关联标签预测，得到业务用户与目标物品之间的预测关联标签；在模型应用推理框架下调用预测模块中的概率预测模块基于解码输入信息的解码表示，对业务用户与目标物品进行关联概率预测，得到预测关联标签下的多个候选概率；在多个候选概率中选择业务用户与目标物品之间的关联度，具体可以在多个候选概率中选择最大的候选概率作为业务用户与目标物品之间的关联度。

除了在物品处理模型的推理过程中对物品处理模型的结构进行改造，本申请实施例还对内存缓存进行优化，对提示信息中出现过的业务用户的信息、关联物品的特征以及候选物品的特征进行表征缓存，表征缓存是指将出现过的业务用户的信息、关联物品的特征以及候选物品的特征中单词的表征信息进行缓存，当出现重复的单词时，可以直接使用缓存的表征信息，而不用再次进行表征分析，减少物品处理模型中编码器的编码耗时。也就是说，本申请实施例在物品处理模型的编码器之前增加缓存模块，对于物品推荐业务，物料库是优先的，不同的业务用户多次请求中的业务用户的信息、关联用户的特征和候选物品的特征可能是重复的，因此，可以将物品特征和业务用户的信息提前计算出表征信息，减少实时表征分析的计算时间。

将内存缓存改造映射到目标物品的提示信息包括的各个单词中，获取目标物品的提示信息中各个单词的表征信息的过程，可以包括：可以获取表征信息库（即缓存模块），表征信息库中可以包括多个词汇（即出现过的单词）的表征信息；可以在表征信息库中查找目标物品的提示信息中各个单词的表征信息；若在表征信息库中查找到与目标物品的提示信息中的第一单词相匹配的词汇，则可以将相匹配的词汇的表征信息，确定为第一单词的表征信息；对目标物品的提示信息中除第一单词外的第二单词进行表征分析，得到第二单词的表征信息。也就是说，在目标物品的提示信息包括的各个单词中，对于已经进行过表征分析的单词，其表征信息缓存在缓存模块中，可以直接在缓存模块中获取，未进行表征分析的单词其表征信息在缓存模块中不存在，需要进行表征分析。

正如前文所描述的，在物品处理模型正式上线进行物品推荐之前，物品处理模型需要经历模型训练阶段和模型测试阶段，在物品处理模型通过模型测试后，可以正式上线进行物品推荐。下面详细介绍物品处理模型的模型测试过程：样本数据中还可以包括用于对训练得到的物品处理模型进行测试的测试集，测试集可以包括多个测试样本，每个测试

样本可以包括测试物品的提示信息和测试物品对应的测试关联标签；可以在模型训练推理框架下基于测试集对物品处理模型的模型性能进行评估，得到第一评估信息，在模型应用推理框架下基于测试集对物品处理模型的模型性能进行评估，得到第二评估信息；若第一评估信息和第二评估信息之间的差异小于差异阈值，则可以确定物品处理模型测试通过，可以在模型应用推理框架下调用物品处理模型对业务用户与为业务用户召回的每个候选物品进行关联预测；若第一评估信息和第二评估信息之间的差异大于或等于差异阈值，则可以确定物品处理模型测试未通过，需要继续对物品处理模型进行调整。

其中，评估信息（即第一评估信息和第二评估信息）具体可以是指预估的 AUC 值，在模型训练推理框架下基于测试集对物品处理模型的模型性能进行评估的过程，与在模型应用推理框架下基于测试集对物品处理模型的模型性能进行评估的过程类似，在此介绍在模型训练推理框架下基于测试集对物品处理模型的模型性能进行评估的过程，具体可以包括：测试集中的测试样本可以包括正测试样本和负测试样本，正测试样本是指测试用户愿意访问的测试物品，负测试样本是指测试用户拒绝访问的测试物品；可以在模型训练推理框架下调用物品处理模型对测试集中的测试样本与相应测试用户进行关联预测，得到预测关联标签；可以统计正测试样本中被正确判断为阳性的第一样本比例，第一样本比例具体可以是正测试样本中被正确判断为阳性的正测试样本与所有的正测试样本之间的比值；可以统计负测试样本中被错误判断为阳性的第二样本比例，第二样本比例具体可以是负测试样本中被错误判断为阳性的负测试样本与所有的负测试样本之间的比值；然后，可以根据第一样本比例和第二样本比例之间的比值，确定第一评估信息；其中，阳性是指预测关联标签指示测试用户愿意访问对应的测试物品。

S303，基于业务用户与目标物品之间的关联度，对目标物品进行业务处理。

本实施例中，可根据业务用户与目标物品之间的关联度对目标物品进行分类，从而针对业务用户对分类到不同类别的目标物品进行不同的业务处理。业务处理是针对业务用户并基于目标物品所做的特定业务下的处理，具体可以是：推荐目标物品，决定不推荐目标物品、将目标物品与业务用户绑定或者将目标物品与业务用户解除绑定关系等。

在业务（即物品推荐业务）中，当业务用户访问业务时，业务服务方可以为业务用户召回多个候选物品，可以在模型应用推理框架下调用物品处理模型对业务用户与每个候选物品进行关联预测，得到业务用户与每个候选物品之间的预测关联标签和预测关联标签下的关联度；其中，任一个候选物品的预测关联标签可以用于指示业务用户愿意访问对应的候选物品，或者，业务用户拒绝访问对应的候选物品。进一步地，可以在为业务用户召回的候选物品中，在多个候选物品中选择预测关联标签用于指示业务用户愿意访问的意向候选物品，然后可以基于意向候选物品的关联度对意向候选物品进行排序，可以将意向候选物品中排列于目标位置之前的意向候选物品作为业务用户的推荐物品，从而可以将推荐物品推荐给业务用户。

也就是说，针对目标物品，基于业务用户与目标物品之间的关联度，对目标物品进行业务处理，可以包括：若目标物品对应的预测关联标签指示业务用户愿意访问目标物品，则可以检查目标物品在意向候选物品中的排列位置，若目标物品在意向候选物品中的排列位置在目标位置之前，则可以将目标物品作为推荐物品推荐给业务用户；若目标物品对应的预测关联标签指示业务用户拒绝访问目标物品，则可以拒绝向业务用户推荐目标物品；若目标物品在意向候选物品中的排列位置为目标位置或在目标位置之后，则可以拒绝向业务用户推荐目标物品。

本申请实施例中，通过对原本为大规模生成式语言模型的物品处理模型进行适应性改造，使得物品处理模型可以改造为适应物品推荐业务的分类模型，使得大规模生成式语言模型在物品推荐业务中的落地称为可能。并且，通过对物品处理模型的训练阶段的推理框架和应用阶段的推理框架进行分离，在物品处理模型的推理阶段使用能够使得物品处理模型的处理速度提升的推理加速框架，可以提升物品推荐效率，从而提升业务处理效率。另

外，在物品处理模型的推理过程中，通过缓存提示信息中已出现单词的表征信息，可以在重复出现单词时直接获取其表征信息，而不用花费时间进行表征分析，进一步提升了模型推理速度，提升物品推荐效率，从而提升业务处理效率。

本申请实施例提供一种业务处理方法，该业务处理方法介绍模型应用推理框架对物品处理模型的加速方式，以及物品处理模型的并发机制。该业务处理方法可以由计算机设备执行，计算机设备例如可以是图 1 所示的业务处理系统中的物品推荐设备 102，或者，计算机设备例如可以是推荐请求设备 101 和物品推荐设备 102 的集成设备。如图 7 所示，该业务处理方法可以包括但不限于以下步骤 S701-步骤 S704：

S701，获取为业务用户召回的目标物品的提示信息，目标物品的提示信息包括业务用户的信息、与业务用户关联的至少一个关联物品的特征、以及目标物品的特征。

本申请实施例中步骤 S701 的执行过程与上述图 3 所示实施例中步骤 S301 的执行过程相同，具体执行过程可以参见上述图 3 所示实施例中步骤 S301 的描述，在此不再赘述。

S702，按照模型应用推理框架定义的优化方式对物品处理模型进行优化处理。

S703，调用优化处理后的物品处理模型基于目标物品的提示信息，对业务用户与目标物品进行关联预测，得到业务用户与目标物品之间的关联度。

在步骤 S702-步骤 S703 中，正如前文所描述的，物品处理模型在模型应用推理框架中的处理速度高于物品处理模型在模型训练推理框架中的处理速度，这是因为模型应用推理框架定义了物品处理模型进行关联预测的优化方式，可以按照模型应用推理框架定义的优化方式对物品处理模型进行优化处理，然后，调用优化处理后的物品处理模型基于目标物品的提示信息，对业务用户与目标物品进行关联预测，得到业务用户与目标物品之间的关联度，这样可以提升物品处理模型进行关联预测的效率。模型应用推理框架的优化方式主要体现在算子融合和矩阵计算优化中的任意一种或两种，下面分别介绍这两种模型应用推理框架的优化方式：

（1）算子融合：

物品处理模型的关联预测逻辑可以由物品处理模型中的各层算子之间相互配合实现的，每层算子需要调用计算机设备 CPU（Central Processing Unit，中央处理器）的 Kernel（内核）进行运算，如果物品处理模型中算子的层数过多，调用内核以及内核之间的交互需要消耗大量的时间，因此，可以对物品处理模型中的各层算子进行算子融合。具体来说，模型应用推理框架的优化方式可以包括物品处理模型中各层算子之间的融合方式，在此情况下，按照优化方式对物品处理模型进行优化处理，可以包括：按照融合方式对物品处理模型中的各层算子进行融合处理；调用优化处理后的物品处理模型基于目标物品的提示信息，对业务用户与目标物品进行关联预测，得到业务用户与目标物品之间的关联度，可以包括：调用融合处理后的物品处理模型基于目标物品的提示信息，对业务用户与目标物品进行关联预测，得到业务用户与目标物品之间的关联度。

进一步地，物品处理模型中的各层算子可以包括矩阵运算算子和非矩阵运算算子，而矩阵运算是物品处理模型进行关联预测的核心运算，因此矩阵运算算子需要消耗大量的时间进行矩阵运算。在此情况下，按照融合方式对物品处理模型中的各层算子进行融合处理，可以包括：保留矩阵运算算子不变，将非矩阵运算算子进行融合处理。

也就是说，本申请实施例可以将物品处理模型中一些运算耗时较短的算子进行融合处理，减少这些算子分别调用内核以及内核之间交互消耗的时间，从而提升物品处理模型的关联预测效率，这样可以提升物品推荐效率，从而提升业务处理效率。

（2）矩阵计算优化：

正如前文所描述的，矩阵运算是物品处理模型进行关联预测的核心运算，矩阵运算在物品处理模型的关联预测过程中耗时较长，因此，优化矩阵运算算法，提升矩阵运算速度，可以提升物品处理模型的关联预测效率。模型应用推理框架的优化方式可以包括物品处理

模型中矩阵运算算子的矩阵运算优化方式,在此情况下,按照优化方式对物品处理模型进行优化处理,可以包括:按照矩阵运算优化方式,对矩阵运算算子的矩阵运算算法进行优化处理。也就是说,本申请实施例可以优化物品处理模型中的矩阵运算算子的矩阵运算算法,这样可以提升矩阵运算算子的矩阵运算速度,提升物品推荐效率,从而提升业务处理效率。

S704,基于业务用户与目标物品之间的关联度,对目标物品进行业务处理。

本申请实施例中步骤**S704**的执行过程与上述图3所示实施例中步骤**S303**的执行过程相同,具体执行过程可以参见上述图3所示实施例中步骤**S303**的描述,在此不再赘述。

在上述步骤**S701**-步骤**S704**的基础上,为了进一步提升物品推荐效率,本申请实施例可以支持并发机制,本申请实施例支持的并发机制可以包括第一并发机制、第二并发机制以及第三并发机制中的任一种;其中,第一并发机制是指对单个业务用户与不同的候选物品并发进行关联预测,第二并发机制是指对单个业务用户与不同业务场景下的候选物品并发进行关联预测,第三并发机制是指对多个业务用户并发进行关联预测。下面对本申请实施例支持的不同并发机制进行介绍:

(1) 第一并发机制:

如图8所示,目标物品是为业务用户召回的多个候选物品中的任一个,多个候选物品可以被划分为N个候选物品集合(如图8所示的候选物品集1-候选物品集N),N为大于1的整数。在此情况下,可以在模型应用推理框架下调用物品处理模型依次对N个候选物品集合中的每个候选物品集合进行关联预测;针对每个候选物品集合,可以在模型应用推理框架下调用物品处理模型,对业务用户与候选物品集合中的每个候选物品并发进行关联预测,得到业务用户与候选物品集合中的每个候选物品之间的关联度,也就是说,每个候选物品集合中的各个候选物品的关联预测过程是并发的。进一步可以理解为,业务用户和每个候选物品集合中的候选物品可以分别组成一个请求包,一共组成了N个“业务用户-候选物品集”的请求包,每个请求包并行请求物品处理模型,物品处理模型并行处理请求包中的多个候选物品。

通过第一并发机制,可以并发对单个业务用户与不同的候选物品进行关联预测,而不是依次对单个业务用户与不同的候选物品进行关联预测,这样可以提升对单个业务用户进行物品推荐的效率,从而提升单个业务用户的业务处理效率。

(2) 第二并发机制:

如图9所示,目标物品是为业务用户召回的多个候选物品中的任一个;多个候选物品属于同一业务的N个不同业务场景(如图9所示的业务场景1-业务场景N),N个业务场景包括的候选物品数量可以是相同的或不同的,N为大于1的整数;在模型应用推理框架下部署了N个物品处理模型,N个物品处理模型与N个业务场景一一对应。在此情况下,可以在模型应用推理框架下调用N个物品处理模型基于相应业务场景中的每个候选物品的提示信息,对业务用户与相应业务场景中的每个候选物品进行关联预测,得到业务用户与相应业务场景中的每个候选物品之间的关联度;其中,N个物品处理模型的关联预测过程并行。也就是说,业务用户和每个业务场景中的候选物品可以分别组成一个请求包,一共组成了N个“业务用户-业务场景下的候选物品”的请求包,并行请求N个物品处理模型,N个物品处理模型并行处理N个请求包。

需要注意的是,N个物品处理模型可以部署于同一设备中或N个不同的设备中。当N个物品处理模型部署于同一设备中时,可以理解为,每个物品处理模型作为设备的一个物品处理服务,N个物品处理服务并行处理N个请求包。当N个物品处理模型部署于N个不同的设备中时,可以理解为,计算机设备是由N个设备组成的分布式系统,N个设备分布式并行处理N个请求包。

通过第二并发机制,可以并发对单个业务用户与不同业务场景下的候选物品进行关联预测,而不是依次对单个业务用户与每个业务场景下的候选物品进行关联预测,这样可以

提升对单个业务用户同时进行多个业务场景下的物品推荐的效率，从而提升单个业务用户的业务处理效率。

(3) 第三并发机制：

如图 10 所示，业务服务方为 N 个业务用户（如图 10 所示的业务用户 1-业务用户 N ）分别召回了候选物品， N 为大于 1 的整数；在模型应用推理框架下部署了 N 个物品处理模型， N 个物品处理模型与 N 个业务用户一一对应。在此情况下，可以在模型应用推理框架下调用 N 个物品处理模型对其相应的业务用户进行关联预测；其中， N 个物品处理模型的关联预测过程并行。也就是说， N 个业务用户和每个业务用户对应的候选物品分别组成一个请求包，一共组成了 N 个“业务用户-候选物品”的请求包，并行请求 N 个物品处理模型， N 个物品处理模型并行处理 N 个请求包。

需要注意的是， N 个物品处理模型可以部署于同一设备中或 N 个不同的设备中。当 N 个物品处理模型部署于同一设备中时，可以理解为，每个物品处理模型作为设备的一个物品处理服务， N 个物品处理服务并行处理 N 个请求包。当 N 个物品处理模型部署于 N 个不同的设备中时，可以理解为，计算机设备是由 N 个设备组成的分布式系统， N 个设备分布式并行处理 N 个请求包。

通过第三并发机制，可以并发对多个业务用户进行物品推荐，而不是依次对每个业务用户进行物品推荐，这样可以支持同时接入更多的业务用户进行物品推荐，提升业务服务方对多个业务用户进行物品推荐的效率。

本申请实施例中，模型应用推理框架定义了物品处理模型进行关联预测的优化方式（包括算子融合和矩阵计算优化），按照模型应用推理框架定义的优化方式对物品处理模型进行优化处理后，调用优化处理后的物品处理模型进行关联预测，可以提升物品处理模型进行关联预测的效率，提升物品推荐效率，从而可以提升业务处理效率。另外，基于并发机制，本申请实施例可以并发对单个业务用户与不同的候选物品进行关联预测，提升对单个业务用户进行物品推荐的效率；基于并发机制，本申请实施例也可以并发对单个业务用户与不同业务场景下的候选物品进行关联预测，以提升对单个业务用户同时进行多个业务场景下的物品推荐的效率；基于并发机制，本申请实施例也可以并发对多个业务用户进行物品推荐，提升业务服务方同时对多个业务用户进行物品推荐的效率。

上述详细阐述了本申请实施例的方法，为了便于更好地实施本申请实施例的上述方案，相应地，下面提供了本申请实施例的装置。

请参见图 11，图 11 是本申请实施例提供的一种业务处理装置的结构示意图，该业务处理装置可以设置于本申请实施例提供的计算机设备中，计算机设备例如可以是图 1 所示的业务处理系统中的物品推荐设备 102，或者可以是图 1 所示的业务处理系统中的推荐请求设备 101 与物品推荐设备 102 的集成设备。图 11 所示的业务处理装置可以是运行于计算机设备中的一个计算机程序，该业务处理装置可以用于执行图 3 或图 7 所示的方法实施例中的部分或全部步骤。请参见图 11，该业务处理装置可以包括如下单元：

获取单元 1101，用于获取为业务用户召回的目标物品的提示信息，目标物品的提示信息包括业务用户的信息、与业务用户关联的至少一个关联物品的特征、以及目标物品的特征；

处理单元 1102，用于在模型应用推理框架下调用物品处理模型，基于目标物品的提示信息预测业务用户与目标物品之间的关联度；物品处理模型在模型训练推理框架下训练得到，并被复用至模型应用推理框架中；物品处理模型在模型应用推理框架中的处理速度高于物品处理模型在模型训练推理框架中的处理速度；

处理单元 1102，还用于基于业务用户与目标物品之间的关联度，对目标物品进行业务处理。

在一种实现方式中，模型应用推理框架定义了物品处理模型进行关联预测的优化方式；

处理单元 1102, 用于在模型应用推理框架下调用物品处理模型基于目标物品的提示信息, 对业务用户与目标物品进行关联预测, 得到业务用户与目标物品之间的关联度时, 具体用于执行如下步骤:

按照优化方式对物品处理模型进行优化处理;

调用优化处理后的物品处理模型基于目标物品的提示信息, 对业务用户与目标物品进行关联预测, 得到业务用户与目标物品之间的关联度。

在一种实现方式中, 优化方式包括物品处理模型中各层算子之间的融合方式;

处理单元 1102, 用于按照优化方式对物品处理模型进行优化处理时, 具体用于执行如下步骤:

按照融合方式对物品处理模型的各层算子进行融合处理;

处理单元 1102, 用于调用优化处理后的物品处理模型基于目标物品的提示信息, 对业务用户与目标物品进行关联预测, 得到业务用户与目标物品之间的关联度时, 具体用于执行如下步骤: 调用融合处理后的物品处理模型基于目标物品的提示信息, 对业务用户与目标物品进行关联预测, 得到业务用户与目标物品之间的关联度。

在一种实现方式中, 优化方式包括物品处理模型中矩阵运算算子的矩阵运算优化方式;

处理单元 1102, 用于按照优化方式对物品处理模型进行优化处理时, 具体用于执行如下步骤:

按照矩阵运算优化方式, 述矩阵运算算子的矩阵运算算法进行优化处理。

在一种实现方式中, 物品处理模型包括编码器和解码器, 解码器包括解码模块和预测模块; 目标物品的提示信息在输入物品处理模型后被划分为多个单词; 处理单元 1102, 用于在模型应用推理框架下调用物品处理模型, 基于目标物品的提示信息预测业务用户与目标物品之间的关联度时, 具体用于执行如下步骤:

获取目标物品的提示信息中各个单词的表征信息; 在模型应用推理框架下调用编码器基于各个单词的表征信息之间的关联性, 对每个单词进行语义分析, 得到每个单词的上下文语义信息; 在模型应用推理框架下调用解码模块基于每个单词的上下文语义信息与解码输入信息的上下文语义信息之间的关联性, 生成解码输入信息的解码表示; 在模型应用推理框架下调用预测模块基于解码输入信息的解码表示, 对业务用户与目标物品进行关联预测, 得到业务用户与目标物品之间的关联度。

在一种实现方式中, 预测模块是采用概率预测模块对标签预测模块进行改造得到的; 处理单元 1102, 用于在模型应用推理框架下调用预测模块基于解码输入信息的解码表示, 对业务用户与目标物品进行关联预测, 得到业务用户与目标物品之间的关联度时, 具体用于执行如下步骤:

在模型应用推理框架下调用标签预测模块基于解码输入信息的解码表示, 对业务用户与目标物品进行关联标签预测, 得到业务用户与目标物品之间的预测关联标签; 在模型应用推理框架下调用概率预测模块基于解码输入信息的解码表示, 对业务用户与目标物品进行关联概率预测, 得到预测关联标签下的多个候选概率;

在多个候选概率中选择业务用户与目标物品之间的关联度。

在一种实现方式中, 处理单元 1102, 用于获取目标物品的提示信息中各个单词的表征信息时, 具体用于执行如下步骤: 获取表征信息库, 表征信息库中包括多个词汇的表征信息; 在表征信息库中查找目标物品的提示信息中各个单词的表征信息; 若在表征信息库中查找到与目标物品的提示信息中的第一单词相匹配的词汇, 则将相匹配的词汇的表征信息, 确定为第一单词的表征信息; 对目标物品的提示信息中除第一单词外的第二单词进行表征分析, 得到第二单词的表征信息。

在一种实现方式中, 目标物品是为业务用户召回的多个候选物品中的任一个; 多个候选物品属于同一业务的 N 个不同业务场景, N 为大于 1 的整数; 在模型应用推理框架下部署了 N 个物品处理模型, N 个物品处理模型与 N 个业务场景一一对应;

处理单元 1102, 还用于执行如下步骤: 在模型应用推理框架下调用 N 个物品处理模型基于相应业务场景中的每个候选物品的提示信息, 对业务用户与相应业务场景中的每个候选物品进行关联预测, 得到业务用户与相应业务场景中的每个候选物品之间的关联度; 其中, N 个物品处理模型的关联预测过程并行。

在一种实现方式中, 处理单元 1102, 还用于执行如下步骤: 获取初始处理模型和样本数据, 样本数据包括训练集和测试集; 在模型训练推理框架下基于训练集对初始处理模型进行训练, 得到物品处理模型; 在模型训练推理框架下基于测试集对物品处理模型的模型性能进行评估, 得到第一评估信息; 在模型应用推理框架下基于测试集对物品处理模型的模型性能进行评估, 得到第二评估信息; 若第一评估信息和第二评估信息之间的差异小于差异阈值, 则在模型应用推理框架下调用物品处理模型对业务用户与为业务用户召回的每个候选物品进行关联预测, 目标物品是为业务用户召回的任一个候选物品。

在一种实现方式中, 测试集中的测试样本包括正测试样本和负测试样本; 正测试样本是指测试用户愿意访问的测试物品, 负测试样本是指测试用户拒绝访问的测试物品; 处理单元 1102, 用于在模型训练推理框架下基于测试集对物品处理模型的模型性能进行评估, 得到第一评估信息时, 具体用于执行如下步骤:

在模型训练推理框架下调用物品处理模型对测试集中的测试样本与相应测试用户进行关联预测, 得到预测关联标签; 统计正测试样本中被正确判断为阳性的第一样本比例; 统计负测试样本中被错误判断为阳性的第二样本比例; 根据第一样本比例和第二样本比例之间的比值, 确定第一评估信息; 其中, 阳性是指预测关联标签指示测试用户愿意访问对应的测试物品。

根据本申请的另一个实施例, 图 11 所示的业务处理装置中的各个单元可以分别或全部合并为一个或若干个另外的单元来构成, 或者其中的某个(些)单元还可以再拆分为功能上更小的多个单元来构成, 这可以实现同样的操作, 而不影响本申请的实施例的技术效果的实现。上述单元是基于逻辑功能划分的, 在实际应用中, 一个单元的功能也可以由多个单元来实现, 或者多个单元的功能由一个单元实现。在本申请的其它实施例中, 业务处理装置也可以包括其它单元, 在实际应用中, 这些功能也可以由其它单元协助实现, 并且可以由多个单元协作实现。

根据本申请的另一个实施例, 可以通过在包括中央处理单元(CPU)、随机存取存储介质(RAM)、只读存储介质(ROM)等处理元件和存储元件的例如计算机的通用计算设备上运行能够执行如图 3 或图 7 所示的部分或全部方法所涉及的各步骤的计算机程序, 来构造如图 11 中所示的业务处理装置, 以及来实现本申请实施例的业务处理方法。计算机程序可以记载于例如计算机可读存储介质上, 并通过计算机可读存储介质装载于上述计算设备中, 并在其中运行。

本申请实施例中, 目标物品是为业务用户召回的, 在获取到为业务用户召回的目标物品的提示信息后, 可以在模型应用推理框架下调用物品处理模型基于目标物品的提示信息, 对业务用户与目标物品进行关联检测, 得到业务用户与目标物品之间的关联度, 可以基于业务用户与目标物品之间的关联度, 对目标物品进行业务处理; 其中, 物品处理模型可以是在模型训练推理框架下训练得到, 并被复用至模型应用推理框架中, 物品处理模型在模型应用推理框架中的处理速度高于物品处理模型在模型训练推理框架中的处理速度。不难看出, 物品处理模型可以用于在业务中进行业务处理, 模型应用推理框架可以理解为是物品处理模型在模型推理过程中采用的推理框架, 模型训练推理框架可以理解为是物品处理模型在模型训练过程中采用的训练框架, 本申请实施例可以对物品处理模型的训练框架和推理框架进行分离, 在物品处理模型的推理阶段使用能够使得物品处理模型的处理速度提升的推理加速框架, 从而可以提升业务处理效率。

基于上述方法以及装置实施例, 本申请实施例提供了一种计算机设备。请参见图 12, 图 12 是本申请实施例提供的一种计算机设备的结构示意图。图 12 所示的计算机设备至少

包括处理器 1201、输入接口 1202、输出接口 1203 以及计算机可读存储介质 1204。其中，处理器 1201、输入接口 1202、输出接口 1203 以及计算机可读存储介质 1204 可通过总线或其他方式连接。

计算机可读存储介质 1204 可以存储在计算机设备的存储器中，计算机可读存储介质 1204 用于存储计算机程序，计算机程序包括计算机指令，处理器 1201 用于执行计算机可读存储介质 1204 存储的计算机程序。处理器 1201（或称 CPU（Central Processing Unit，中央处理器））是计算机设备的计算核心以及控制核心，其适于实现计算机程序，具体适于加载并执行计算机程序从而实现相应方法流程或相应功能。

本申请实施例还提供了一种计算机可读存储介质（Memory），计算机可读存储介质是计算机设备中的记忆设备，用于存放程序和数据。可以理解的是，此处的计算机可读存储介质既可以包括计算机设备中的内置存储介质，当然也可以包括计算机设备支持的扩展存储介质。计算机可读存储介质提供存储空间，该存储空间存储了计算机设备的操作系统。并且，在该存储空间中还存放了适于被处理器加载并执行的计算机程序。需要说明的是，此处的计算机可读存储介质可以是高速 RAM 存储器，也可以是非不稳定的存储器（Non-Volatile Memory），例如至少一个磁盘存储器；可选的还可以是至少一个位于远离前述处理器的计算机可读存储介质。

计算机设备例如可以是图 1 所示的业务处理系统中的物品推荐设备 102，或者可以是图 1 所示的业务处理系统中的推荐请求设备 101 和物品推荐设备 102 的集成设备，具体实现中，可由处理器 1201 加载并执行计算机可读存储介质 1204 中存放的计算机程序，以实现上述有关图 3 或图 7 所示的业务处理方法中的相应步骤。具体实现中，计算机可读存储介质 1204 中的计算机程序由处理器 1201 加载并执行如下步骤：

获取为业务用户召回的目标物品的提示信息，目标物品的提示信息包括业务用户的信息、与业务用户关联的至少一个关联物品的特征、以及目标物品的特征；

在模型应用推理框架下调用物品处理模型，基于目标物品的提示信息预测业务用户与目标物品之间的关联度；物品处理模型在模型训练推理框架下训练得到，并被复用至模型应用推理框架中；物品处理模型在模型应用推理框架中的处理速度高于物品处理模型在模型训练推理框架中的处理速度；

基于业务用户与目标物品之间的关联度，对目标物品进行业务处理。

在一种实现方式中，模型应用推理框架定义了物品处理模型进行关联预测的优化方式；计算机可读存储介质 1204 中的计算机程序由处理器 1201 加载并执行在模型应用推理框架下调用物品处理模型，基于目标物品的提示信息预测业务用户与目标物品之间的关联度时，具体用于执行如下步骤：按照优化方式对物品处理模型进行优化处理；调用优化处理后的物品处理模型基于目标物品的提示信息，对业务用户与目标物品进行关联预测，得到业务用户与目标物品之间的关联度。

在一种实现方式中，优化方式包括物品处理模型中各层算子之间的融合方式；

计算机可读存储介质 1204 中的计算机程序由处理器 1201 加载并执行按照优化方式对物品处理模型进行优化处理时，具体用于执行如下步骤：按照融合方式对物品处理模型各层算子进行融合处理；

计算机可读存储介质 1204 中的计算机程序由处理器 1201 加载并执行调用优化处理后的物品处理模型基于目标物品的提示信息，对业务用户与目标物品进行关联预测，得到业务用户与目标物品之间的关联度时，具体用于执行如下步骤：

调用融合处理后的物品处理模型基于目标物品的提示信息，对业务用户与目标物品进行关联预测，得到业务用户与目标物品之间的关联度。

在一种实现方式中，优化方式包括物品处理模型中矩阵运算算子的矩阵运算优化方式；

计算机可读存储介质 1204 中的计算机程序由处理器 1201 加载并执行按照优化方式对物品处理模型进行优化处理时，具体用于执行如下步骤：

按照矩阵运算优化方式，述矩阵运算算子的矩阵运算算法进行优化处理。

在一种实现方式中，物品处理模型包括编码器和解码器，解码器包括解码模块和预测模块；目标物品的提示信息在输入物品处理模型后被划分为多个单词；计算机可读存储介质 1204 中的计算机程序由处理器 1201 加载并执行在模型应用推理框架下调用物品处理模型，基于目标物品的提示信息预测业务用户与目标物品之间的关联度时，具体用于执行如下步骤：

获取目标物品的提示信息中各个单词的表征信息；在模型应用推理框架下调用编码器基于各个单词的表征信息之间的关联性，对每个单词进行语义分析，得到每个单词的上下文语义信息；在模型应用推理框架下调用解码模块基于每个单词的上下文语义信息与解码输入信息的上下文语义信息之间的关联性，生成解码输入信息的解码表示；在模型应用推理框架下调用预测模块基于解码输入信息的解码表示，对业务用户与目标物品进行关联预测，得到业务用户与目标物品之间的关联度。

在一种实现方式中，预测模块是采用概率预测模块对标签预测模块进行改造得到的；计算机可读存储介质 1204 中的计算机程序由处理器 1201 加载并执行在模型应用推理框架下调用预测模块基于解码输入信息的解码表示，对业务用户与目标物品进行关联预测，得到业务用户与目标物品之间的关联度时，具体用于执行如下步骤：

在模型应用推理框架下调用标签预测模块基于解码输入信息的解码表示，对业务用户与目标物品进行关联标签预测，得到业务用户与目标物品之间的预测关联标签；在模型应用推理框架下调用概率预测模块基于解码输入信息的解码表示，对业务用户与目标物品进行关联概率预测，得到预测关联标签下的多个候选概率；在多个候选概率中选择业务用户与目标物品之间的关联度。

在一种实现方式中，计算机可读存储介质 1204 中的计算机程序由处理器 1201 加载并执行获取目标物品的提示信息中各个单词的表征信息时，具体用于执行如下步骤：

获取表征信息库，表征信息库中包括多个词汇的表征信息；在表征信息库中查找目标物品的提示信息中各个单词的表征信息；若在表征信息库中查找到与目标物品的提示信息中的第一单词相匹配的词汇，则将相匹配的词汇的表征信息，确定为第一单词的表征信息；对目标物品的提示信息中除第一单词外的第二单词进行表征分析，得到第二单词的表征信息。

在一种实现方式中，目标物品是为业务用户召回的多个候选物品中的任一个；多个候选物品属于同一业务的 N 个不同业务场景， N 为大于 1 的整数；在模型应用推理框架下部署了 N 个物品处理模型， N 个物品处理模型与 N 个业务场景一一对应；

计算机可读存储介质 1204 中的计算机程序由处理器 1201 加载并还用于执行如下步骤：

在模型应用推理框架下调用 N 个物品处理模型基于相应业务场景中的每个候选物品的提示信息，对业务用户与相应业务场景中的每个候选物品进行关联预测，得到业务用户与相应业务场景中的每个候选物品之间的关联度；其中， N 个物品处理模型的关联预测过程并行。

在一种实现方式中，计算机可读存储介质 1204 中的计算机程序由处理器 1201 加载并还用于执行如下步骤：获取初始处理模型和样本数据，样本数据包括训练集和测试集；在模型训练推理框架下基于训练集对初始处理模型进行训练，得到物品处理模型；在模型训练推理框架下基于测试集对物品处理模型的模型性能进行评估，得到第一评估信息；在模型应用推理框架下基于测试集对物品处理模型的模型性能进行评估，得到第二评估信息；若第一评估信息和第二评估信息之间的差异小于差异阈值，则在模型应用推理框架下调用物品处理模型对业务用户与为业务用户召回的每个候选物品进行关联预测，目标物品是为业务用户召回的任一个候选物品。

在一种实现方式中，测试集中的测试样本包括正测试样本和负测试样本；正测试样本

是指测试用户愿意访问的测试物品，负测试样本是指测试用户拒绝访问的测试物品；计算机可读存储介质 1204 中的计算机程序由处理器 1201 加载并执行在模型训练推理框架下基于测试集对物品处理模型的模型性能进行评估，得到第一评估信息时，具体用于执行如下步骤：在模型训练推理框架下调用物品处理模型对测试集中的测试样本与相应测试用户进行关联预测，得到预测关联标签；统计正测试样本中被正确判断为阳性的第一样本比例；统计负测试样本中被错误判断为阳性的第二样本比例；根据第一样本比例和第二样本比例之间的比值，确定第一评估信息；其中，阳性是指预测关联标签指示测试用户愿意访问对应的测试物品。

本申请实施例中，目标物品是为业务用户召回的，在获取到为业务用户召回的目标物品的提示信息后，可以在模型应用推理框架下调用物品处理模型，基于目标物品的提示信息预测业务用户与目标物品之间的关联度，可以基于业务用户与目标物品之间的关联度，对目标物品进行业务处理；其中，物品处理模型可以是在模型训练推理框架下训练得到，并被复用至模型应用推理框架中，物品处理模型在模型应用推理框架中的处理速度高于物品处理模型在模型训练推理框架中的处理速度。不难看出，物品处理模型可以在业务中用于进行业务处理，模型应用推理框架可以理解为是物品处理模型在模型推理过程中采用的推理框架，模型训练推理框架可以理解为是物品处理模型在模型训练过程中采用的训练框架，本申请实施例可以对物品处理模型的训练框架和推理框架进行分离，在物品处理模型的推理阶段使用能够使得物品处理模型的处理速度提升的推理加速框架，从而可以提升业务处理效率。

根据本申请的一个方面，提供了一种计算机程序产品，该计算机程序产品包括计算机程序，该计算机程序存储在计算机可读存储介质中。计算机设备的处理器从计算机可读存储介质读取该计算机程序，处理器执行该计算机程序，使得该计算机设备执行上述各种可选方式中提供的业务处理方法。

以上实施例的各技术特征可以进行任意的组合，为使描述简洁，未对上述实施例中的各个技术特征所有可能的组合都进行描述，然而，只要这些技术特征的组合不存在矛盾，都应当认为是本说明书记载的范围。

以上所述实施例仅表达了本申请的几种实施方式，其描述较为具体和详细，但并不能因此而理解为对发明专利范围的限制。应当指出的是，对于本领域的普通技术人员来说，在不脱离本申请构思的前提下，还可以做出若干变形和改进，这些都属于本申请的保护范围。因此，本申请专利的保护范围应以所附权利要求为准。

权利要求

1、一种业务处理方法，由计算机设备执行，包括：

获取为业务用户召回的目标物品的提示信息，所述目标物品的提示信息包括所述业务用户的信息、与所述业务用户关联的至少一个关联物品的特征、以及所述目标物品的特征；

在模型应用推理框架下调用物品处理模型，基于所述目标物品的提示信息预测所述业务用户与所述目标物品之间的关联度；所述物品处理模型在模型训练推理框架下训练得到，并被复用至所述模型应用推理框架中；所述物品处理模型在所述模型应用推理框架中的处理速度高于所述物品处理模型在所述模型训练推理框架中的处理速度；及

基于所述业务用户与所述目标物品之间的关联度，对所述目标物品进行业务处理。

2、如权利要求1所述的方法，所述模型应用推理框架定义了所述物品处理模型进行关联预测的优化方式；所述在模型应用推理框架下调用物品处理模型，基于所述目标物品的提示信息预测所述业务用户与所述目标物品之间的关联度，包括：

按照所述优化方式对所述物品处理模型进行优化处理；

调用优化处理后的所述物品处理模型基于所述目标物品的提示信息，对所述业务用户与所述目标物品进行关联预测，得到所述业务用户与所述目标物品之间的关联度。

3、如权利要求2所述的方法，所述优化方式包括所述物品处理模型中各层算子之间的融合方式；

所述按照所述优化方式对所述物品处理模型进行优化处理，包括：

按照所述融合方式对所述物品处理模型各层的算子进行融合处理；

所述调用优化处理后的所述物品处理模型基于所述目标物品的提示信息，对所述业务用户与所述目标物品进行关联预测，得到所述业务用户与所述目标物品之间的关联度，包括：

调用融合处理后的所述物品处理模型基于所述目标物品的提示信息，对所述业务用户与所述目标物品进行关联预测，得到所述业务用户与所述目标物品之间的关联度。

4、如权利要求2或3所述的方法，所述优化方式包括所述物品处理模型中矩阵运算算子的矩阵运算优化方式；

所述按照所述优化方式对所述物品处理模型进行优化处理，包括：

按照所述矩阵运算优化方式，对所述矩阵运算算子的矩阵运算算法进行优化处理。

5、如权利要求1-4任一项所述的方法，所述物品处理模型包括编码器和解码器，所述解码器包括解码模块和预测模块；所述目标物品的提示信息在输入所述物品处理模型后被划分为多个单词；所述在模型应用推理框架下调用物品处理模型，基于所述目标物品的提示信息预测所述业务用户与所述目标物品之间的关联度，包括：

获取所述目标物品的提示信息中各个单词的表征信息；

在所述模型应用推理框架下调用所述编码器基于各个所述单词的表征信息之间的关联性，对每个所述单词进行语义分析，得到每个所述单词的上下文语义信息；

在所述模型应用推理框架下调用所述解码模块基于每个所述单词的上下文语义信息与解码输入信息的上下文语义信息之间的关联性，生成所述解码输入信息的解码表示；

在所述模型应用推理框架下调用所述预测模块基于所述解码输入信息的解码表示，对所述业务用户与所述目标物品进行关联预测，得到所述业务用户与所述目标物品之间的关联度。

6、如权利要求5所述的方法，所述预测模块是采用概率预测模块对标签预测模块进行改造得到的；所述在所述模型应用推理框架下调用所述预测模块基于所述解码输入信息的解码表示，对所述业务用户与所述目标物品进行关联预测，得到所述业务用户与所述目标物品之间的关联度，包括：

在所述模型应用推理框架下调用所述标签预测模块基于所述解码输入信息的解码表

示, 对所述业务用户与所述目标物品进行关联标签预测, 得到所述业务用户与所述目标物品之间的预测关联标签;

在所述模型应用推理框架下调用所述概率预测模块基于所述解码输入信息的解码表示, 对所述业务用户与所述目标物品进行关联概率预测, 得到所述预测关联标签下的多个候选概率;

在所述多个候选概率中选择所述业务用户与所述目标物品之间的关联度。

7、如权利要求 5 或 6 所述的方法, 所述获取所述目标物品的提示信息中各个单词的表征信息, 包括:

获取表征信息库, 所述表征信息库中包括多个词汇的表征信息;

在所述表征信息库中查找所述目标物品的提示信息中各个单词的表征信息;

若在所述表征信息库中查找到与所述目标物品的提示信息中的第一单词相匹配的词汇, 则将所述相匹配的词汇的表征信息, 确定为所述第一单词的表征信息;

对所述目标物品的提示信息中除所述第一单词外的第二单词进行表征分析, 得到所述第二单词的表征信息。

8、如权利要求 1-7 任一项所述的方法, 所述目标物品是为所述业务用户召回的多个候选物品中的任一个; 所述多个候选物品属于同一业务的 N 个不同业务场景, N 为大于 1 的整数; 在所述模型应用推理框架下部署了 N 个物品处理模型, 所述 N 个物品处理模型与 N 个业务场景一一对应;

所述方法还包括:

在所述模型应用推理框架下调用所述 N 个物品处理模型基于相应业务场景中的每个所述候选物品的提示信息, 对所述业务用户与相应业务场景中的每个所述候选物品进行关联预测, 得到所述业务用户与相应业务场景中的每个所述候选物品之间的关联度;

其中, 所述 N 个物品处理模型的关联预测过程并行。

9、如权利要求 1-8 任一项所述的方法, 所述方法还包括:

获取初始处理模型和样本数据, 所述样本数据包括训练集和测试集;

在所述模型训练推理框架下基于所述训练集对所述初始处理模型进行训练, 得到所述物品处理模型;

在所述模型训练推理框架下基于所述测试集对所述物品处理模型的模型性能进行评估, 得到第一评估信息;

在所述模型应用推理框架下基于所述测试集对所述物品处理模型的模型性能进行评估, 得到第二评估信息;

若所述第一评估信息和所述第二评估信息之间的差异小于差异阈值, 则在所述模型应用推理框架下调用所述物品处理模型对所述业务用户与为所述业务用户召回的每个候选物品进行关联预测, 所述目标物品是为所述业务用户召回的任一个候选物品。

10、如权利要求 9 所述的方法, 所述测试集中的测试样本包括正测试样本和负测试样本; 所述正测试样本是指测试用户愿意访问的测试物品, 所述负测试样本是指测试用户拒绝访问的测试物品; 所述在所述模型训练推理框架下基于所述测试集对所述物品处理模型的模型性能进行评估, 得到第一评估信息, 包括:

在所述模型训练推理框架下调用所述物品处理模型对所述测试集中的测试样本与相应测试用户进行关联预测, 得到预测关联标签;

统计所述正测试样本中被正确判断为阳性的第一样本比例;

统计所述负测试样本中被错误判断为阳性的第二样本比例;

根据所述第一样本比例和所述第二样本比例之间的比值, 确定所述第一评估信息;

其中, 所述阳性是指所述预测关联标签指示测试用户愿意访问对应的测试物品。

11、一种业务处理装置, 包括:

获取单元, 用于获取为业务用户召回的目标物品的提示信息, 所述目标物品的提示信

息包括所述业务用户的信息、与所述业务用户关联的至少一个关联物品的特征、以及所述目标物品的特征；

处理单元，用于在模型应用推理框架下调用物品处理模型，基于所述目标物品的提示信息预测所述业务用户与所述目标物品之间的关联度；所述物品处理模型在模型训练推理框架下训练得到，并被复用至所述模型应用推理框架中；所述物品处理模型在所述模型应用推理框架中的处理速度高于所述物品处理模型在所述模型训练推理框架中的处理速度；及

所述处理单元，还用于基于所述业务用户与所述目标物品之间的关联度，对所述目标物品进行业务处理。

12、一种计算机设备，所述计算机设备包括：

处理器，适于实现计算机程序；

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，所述计算机程序适于由所述处理器加载并执行如权利要求 1-10 任一项所述的业务处理方法。

13、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，所述计算机程序适于由处理器加载并执行如权利要求 1-10 任一项所述的业务处理方法。

14、一种计算机程序产品，所述计算机程序产品包括计算机程序，所述计算机程序被处理器执行时实现如权利要求 1-10 任一项所述的业务处理方法。