专利申请预检索分析报告

客户文号：	我所文号：241224	**代理师**：王毓理
发明创造名称：多阶段算力回收框架		
申请人：上海交通大学		
联系人及联系方式：		
专利申请类型： ☒发明 ☐实用新型 ☐外观设计		
检索依据的申请文件：☒技术交底文件 ☐专利申请文本		
是否发表过期刊论文：☒否 ☐是		
本申请技术发明点	一种多阶段算力回收框架，根据级联式推荐系统架构，在每一处推荐模型做输出结果的缓存。读取缓存，并将读取的缓存结果作为候选物品集合，送入下一处推荐模型。计算不同算力回收策略的算力开销 c_j；然后利用采集的数据训练效果损耗预估装置并对在线算力回收策略决策问题进行问题建模，并对建模的数学问题进行对偶问题求解，得出实现推荐效果基本持平前提下提高资源利用率的策略打分公式，再通过二分求解最优拉格朗日乘子。在在线阶段，将在线到来的用户请求及请求相关信息送入请求过滤装置，将请求分类为初始请求和重复请求，当若该请求为初始请求，则调用完整的推荐链路，包括召回模型、粗排模型、精排模型、召回模型，生成物品推荐；同时，将各模型的输出结果进行缓存；之后结束服务。当该请求为重复请求，则利用离线训得的效果损耗预估装置与策略打分参数，计算不同策略的打分。
检索时间	2024 年 12 月 28 日
检索数据库	中国发明专利公开文献、中国发明专利公告文献、中国实用新型专利文献、PCT 公开文献、专利信息服务平台(CNIPR)、中国专利检索系统文摘数据库(CPRSABS)、中文全文库(CNTXT)、德温特世界专利索引数据库(DWPI)、CNKI 系列数据库
检索关键词	多阶段算力回收框架、智能算力、结果缓存与复用
检索式	申请全文包含（多阶段算力回收框架和/或智能算力和/或结果缓存与复用）

检索方式	☒自行检索 ☐委外检索（委托单位： ）
检索结果	Dcaf: a dynamic computation allocation framework for online serving system Computation resource allocation solution in recommender systems ;CN202310659624.3，针对推荐系统的算力优化的方法和装置;CN202410406540.3 一种检验工艺的自动推荐系统;CN202311452396.9，基于 GP;CN202111376003.1 一种轻量级 B2B 电商平台的推荐系统算法;CN202310912981.6，一种新媒体内容推荐方法以及推荐系统;CN202210564581.6 一种算力自适应的模型异构联邦推荐方法及系统;CN202410332296.0 一种信息推荐方法、装置、电子设备及存储介质;CN202310312091.1 基于区块链技术的分布式 AI 推理系统
检索分析	**1. 关于本申请新颖性的评述：** 基于上述关键词检索到本领域较为近似的一份现有技术（对比文件 1）：CN202310659624.3 针对推荐系统的算力优化的方法和装置 对比文件 1 公开了：一种针对推荐系统的算力优化的方法和装置，方法包括：获取目标推荐请求；响应于目标推荐请求，确定多个节点中的全部或部分节点构成的若干候选链路；基于若干候选链路包括的各节点，对各节点的候选算力档位进行组合得到多个全局算力档位；预估分别采用多个全局算力档位执行目标推荐请求的各算力消耗；预估目标推荐请求分别在多个全局算力档位对应的各价值预估分；根据各算力消耗和各价值预估分，从多个全局算力档位中选择出目标全局算力档位，作为执行目标推荐请求的算力档位。能够在业务效果不变的情况下，降低算力消耗，且具备很高的准确性和时效性。 由上述内容可知，本申请的技术方案与上述对比文件 1 的区别在于：本申请对不同阶段推荐模型的输出进行缓存，再根据本申请设计的算法，对缓存结果再次利用。从而实现请求个性化

	的算力资源分配，以及算力资源的回收利用，达到推荐效果不变，推荐系统资源利用率获得提升。 新颖性结论：本申请与对比文件 1 相比具有新颖性。 **2. 关于本申请创造性的评述；** 相比之下本申请解决的技术问题是：推荐技术是互联网平台的重要基石，它对平台的运营和用户的留存具有深远影响。目前，工业界推荐系统部署了越来越多复杂的推荐模型来提升系统的推荐效果，却鲜少有研究进行系统的效率提升研究。随着算力投入对效果提升的边际效益递减效应越来越显著，本申请应更优先关注系统的资源利用率。然而，虽然现有的级联式推荐系统架构将单条流量上的效率优化到了极致——可以在数百甚至数十毫秒内响应用户的请求并返回结果，它却为多条流量的分配同样的算力资源。因此现有推荐系统在多条用户流量上仍存在效率优化的空间。 经进一步根据上述技术手段进行检索，发现相关领域中公开的参考文献（对比文件 2）如下，也记载了与上述技术手段相近似、技术效果接近的装置/方法，包括：CN202311452396.9 基于 GPU 集群的高维数据推荐系统及方法 其中记载了：一种基于 GPU 集群的高维数据推荐系统及方法，在客户端对数据训练任务中的训练数据进行数据拆分，将生成的对应训练数据的原始数据模块序列发送到 GPU 集群；根据对应训练数据的原始数据模块序列，得到数据训练任务特征，GPU 集群根据数据训练任务特征生成对应 GPU 集群的算力容器；根据客户端上传的对应训练数据的原始数据模块序列的算力占用，在对应 GPU 集群的算力容器包含的 GPU 单元中分配原始数据模块序列中的各个数据模块；根据数据训练任务进行处理，生成数据训练后的数据，并根据数据训练后的数据，生成推荐数据。通过该技术所提供的技术方案，可以提高数据处理的效率。 但本申请与上述对比文件 2 记载的内容相比，依旧存在一定的区别：本申请中效果损耗预估模型采用用户侧特征如性别、年

	龄、国籍等，请求侧特征如距前次请求时间间隔、前次请求曝光数、点击数等，用户历史点击行为等。本申请用 GRU 对用户历史行为进行序列建模，之后将抽取出的特征与其他映射为低维嵌入的特征拼接，送入多任务学习的门控多专家混合模块（MMoE）。MMoE 会输出 4 个值，包括 $l^{ctrl}, l_1^{uplift}, l_2^{uplift}, l_3^{uplift}$。$l_1^{treat} = l_1^{uplift} + l^{ctrl}$，$l_2^{treat} = l_2^{uplift} + l^{ctrl}$，$l_3^{treat} = l_3^{uplift} + l^{ctrl}$，最后，$l^{ctrl}, l_1^{treat}, l_2^{treat}, l_3^{treat}$ 经过激活函数最终得到输出结果——预估的效果损耗值；两者在解决的技术问题和所采用的技术手段，得到的技术效果上相比并不完全相同：本申请对在线算力回收策略决策问题进行问题建模具体为：建模成约束最优化问题。假设共有 N 次广告请求，请求 i 采用算力回收策略 j 时的效果损耗为 $\Delta Q_{i,j}$，系统要求推荐系统损耗不超过阈值 B。策略分配 $x_{i,j} = 1$表示算法决策请求 i 采用算力回收策略 j，则问题建模为 $\min_x \sum_{i,j} x_{i,j} \cdot c_j + \frac{\omega}{2} \|x\|^2$ $s.t. \quad \sum_{i,j} x_{i,j} \cdot \triangle Q_{i,j} \le B, \quad \sum_j x_{i,j} = 1, \quad x_{i,j} \ge 0,$；因此上述对比文件 1 和对比文件 2 的结合并未公开解决区别技术特征的技术启示，即具有一定创造性。 **3 技术方案应用前景分析** 本申请应用于：推荐系统领域，根据发明人提供的行业技术情报，结合事务所对类似技术的代理经验，该技术方案初步评估应用前景，本申请对级联式广告推荐系统架构进行了创新，通过“缓存”与“回收”的方法，减少广告系统中不必要的算力开销。本申请引入了细粒度算力回收策略，并将计算回收决策表述为在线约束优化问题。对优化问题求解，得到在线服务的策略打分公式，只要近在线更新打分参数，即可实现在线近似最优算力回收决策。最终，ComRecycle 可以在保证相同水平的

	推荐效果的同时实现降低算力消耗的目标。综上，本申请提出多阶段智能算力回收框架，对未曝光商品进行“缓存”与“复用”，即缓存先前未曝光的商品，并在适当机会时重新展示，从而实现多条流量的效率优化。该框架可以提高推荐系统的资源利用率，在本申请的实验中，该框架可以减少 20%的算力资源消耗。 **4 分析结论：** 本申请技术方案与上述文献相比具有新颖性和工业实用性，达到/接近当前专利局创造性审查尺度。
方向性建议	详细公开本申请技术方案实现过程中所采用的各个技术手段的细节、技术改进点的可选参数设置以及详细实验数据，供事务所在修改过程中根据补充修改的技术内容及实验数据进一步提高本申请的创造性。以防由于专利局审查员个体对专利法创造性的认知差异导致申请的创造性并错误评估。
检索意见	本申请技术方案与上述文献相比具有新颖性和工业实用性，达到/接近当前专利局创造性审查尺度。
备注	以上初步评价为根据申请人要求，仅基于技术交底书、申请人课题组提供的信息以及事务所初步查询得到的结果，并不代表专利申请经沟通修改后的文本的专利性。国家知识产权局在申请提交后的实质审查中可能由于情势变更、资源、条件的不同而得出与上述评估结论不同的检索报告。

代理机构（盖章）：上海交达专利事务所

日 期：2024 年 12 月