User query with historical purchase ***Ui***: "1. Moroccan Infusion Deep Conditioning Shine Mask, 2.Smooth, Infusing Smoothing Serum, 3. Self Heating One Minute Mask, 4. Virtually Indestructible Haircut Kit."

Cold-start Item ***A:*** Best Hair Conditioner - Tru Moroccan Argan Oil Conditioner - Gain Silky Shiny Hair Instantly With The Absolute Best Hair Conditioner With Argan Oil

Cold-start Item ***B:*** Clump Crusher Mascara, Very Black 800, 0.44 Ounce

Prompt: The user purchased the following beauty products in order: {{{***Ui***}}}. Predict if the user will prefer to purchase product A or B in the next. A is {{{***A***}}} and B is {{{***B***}}}. Answer A or B.

2024谷歌：从冷启动到热推荐，谷歌借助LLM的数据增强，提升内容冷启动效果

SmartMindAI
专注搜索、广告、推荐、大模型和人工智能最新技术，欢迎关注我
已关注

Introduction

大型语言模型（LLMs）在大规模网络数据上的训练，赋予了它们全面的理解能力，展现出卓越的推理和泛化能力，这些能力已推动多个领域的创新，包括创作性写作、交互式对话系统和搜索引擎设计等。鉴于LLMs在深入理解用户需求和内容特性的潜力，我们深入探讨了其在推荐系统中的应用，尤其强调在解决冷启动问题中的应用。推荐系统作为在线平台上推广引人兴趣内容的关键途径，通过解析用户的先前互动来预测和理解其喜好偏好，并以此为基础，建立个性化推荐，指导用户发现符合其兴趣的内容。

在推荐系统领域，基于模型的方法广泛采用用户与内容ID相关的学习嵌入，旨在记录并预测相关性。这种方法依赖于对用户行为的深入理解和内容间的复杂关系，以及通过预测共同模式来构建个性化推荐策略。在这里，传统推荐系统主要基于协同过滤、矩阵分解等方法，以及最新的深度学习模型，旨在强化用户与内容间的交互，通过感知用户的历史行为模式，预判其未来的兴趣点，从而提供精准的个性化推荐。在用户与内容频繁互动的环境下，利用基于ID的嵌入方法来推荐新颖且个性化的内容时，面临着被称为冷启动的核心挑战，主要涉及新内容缺乏充分的曝光和互动数据。为应对这一难题，基于内容的推荐系统尝试通过运用内容的元信息辅助内容表示学习，通过使用元特征转换或其组合替代理由ID指定的内容嵌入，以此从大量互动的内容中推断出对冷启动内容的理解。

近期大模型的突破性进展为改进推荐系统提供了新机遇。现有研究已探讨了将用户查询转换为文本，并利用生成式大型语言模型（LLM）构建无需特定ID的推荐系统的可能性。然而，为了使推荐模型适应不同场景，仍需针对每种场景对大规模预训练模态编码器进行微调--这一过程既复杂又耗时，需要大量的工程投入。此外，按照上述方法要求为每个用户分别提供LLM或大型基础模型服务以获取推荐结果，结果的生成往往延迟显著，通常远超过推荐系统期望的O(100ms)响应时间界限。因此，满足在线推荐服务需要以每秒处理大量查询的性能要求时，这种方式导致的成本过高，难以实现商业应用。

是否能够生成关于冷启动内容的合成用户行为数据，如用户偏好由二元比较（如"可能更倾向于观看教授Andrew Ng的视频"或"可能对Selena Gomez的新剧感兴趣"）来表示？结合这些合成交互数据是否会改善经典推荐模型在冷启动场景下的表现？本文提出的方法旨在不增加服务时的延迟，同时有效的解决了数据稀缺性问题。 1. 引入了利用LLMs在二元比较提示下的能力，推测用户对内容的选择偏好。 2. 将LLM生成的合成偏好与标准推荐任务融入了一致的损失函数框架。 3. 在实际数据集上进行了实验，结果表明，即使使用较小型的模型，这样的合成数据增强方法也能显著提升冷启动场景下的推荐性能。

Related Work

内容服务可能成本高昂，并阻碍其广泛应用；相反，我们提出在训练阶段利用LLM生成的合成示例供传统基于ID的推荐系统使用，此策略在提升服务效率的同时，保持了系统的实用性。

针对冷启动内容推荐问题，即新上传内容在缺乏用户反馈情况下的推荐挑战，我们关注在对这类内容进行推荐时面临的困难。传统推荐系统通过整合额外的辅助信息+与基于ID的推荐嵌入关联来解决信息不足的问题。尽管已有一些通过元学习方式训练冷启动内容嵌入的研究，但这些方法似乎仍缺乏有效处理交互数据较少的内容策略。我们的方法则寻求利用LLM生成冷启动内容的合成训练线索，进而直接从这些合成线索中学习协作嵌入，以提高冷启动内容的推荐精度。

在数据增强策略中，神经网络模型+的训练效率提升已获证实。在推荐系统场景中，借鉴了先前成功案例，CLS4Rec 通过随机内容剪裁、遮蔽和重新排序，创建了用户历史交互的增强视角，显著提升了模型的鲁棒性+和预测准确性。针对用户交互较低的情况，文献@wang2022learning 引入了一种学习-学习管道，旨在增强这类用户的训练数据集，从而提高整个模型的性能。然而，我们的研究是首个致力于使用历史交互数据生成增强训练数据以填补冷启动内容知识缺口的尝试，专门为解决冷启动内容的知识获取难题而设计。

Preliminaries

用户集$\mathcal{U}$包含个体用户u_1至u_G，平台上的热门内容组$\mathcal{I}_{warm}$包含i_1到i_P，而冷启动+内容组$\mathcal{I}_{cold}$则包含了从i_{P+1}到i_{P+N}的内容。这些内容的相同ID关联着可训练的嵌入表示。个性化推荐的基本逻辑是预测用户与每个内容的相容性，进一步从候选内容集中筛选出与其相容性高的内容，形成个性化的推荐列表。 潜在因素模型，作为大规模矩阵分解的有效变体，被Netflix Prize等场景广泛研究。该模型通过相似的潜在因子向量乘积来近似用户与内容的相容性。具体而言，如果用户u和内容i的潜在因子向量分别为$\mathbf{v_u}$和$\mathbf{v_i}$，那么它们的相容性$\hat{y}_{u,i}$将由 $\mathbf{v_u}^T\mathbf{v_i}$ 计算得来。 当前许多推荐框架是潜在因素模型的拓展版本。然而，在冷启动问题中，由于缺乏数据信号以供获取内容i的嵌入$\mathbf{v_i}$，现存方法面临挑战。为了缓解知识缺陷，我们提出使用合成数据+来模拟用户对冷启动内容的可能交互，以此填补数据空白。

LLMs as Data Augmenters

增强数据生成

我们专注于PaLM系列模型，直接使用其生成结果而无需微调。我们遵循将用户交互过的商品描述放入提示中。具体而言，对于训练集中的用户查询U_i，我们采用商品标题来表示历史互动。为了从描述性用户查询推断用户的偏好，我们可以询问用户是否喜欢某个冷启动商品（点对点）或比较两个冷启动商品A和B（成对）。大型语言模型在精确的点对点相关性评估方面有困难，但在成对比较任务上表现更好。因此，我们利用大型语言模型生成冷启动商品间的成对偏好。

Figure 1: Pairwise comparison prompt for a user query.

具体地，我们随机选取一对商品(A, B)，其中$A, B \in \mathcal{I}_{cold}$，并构建如图所示的提示来获取用户对$A$与$B$的偏好。成对比较确保了每次调用大型语言模型都能得到两件冷启动商品间的偏好信号。

成对比较损失

为了在训练过程中整合这一增强信号，我们将冷启动商品对的成对偏好预测作为辅助任务加入到常规推荐任务中。大型语言模型返回的答案被视为pos项，另一项为neg。如果用户u更偏好冷启动商品pos而非neg，我们采用受贝叶斯⁺个性化排名(BPR)损失启发的以下成对损失函数：

$$\mathcal{L}_{aug} = -\sum_{(u,pos,neg)} \ln \sigma(\hat{y}_{u,pos} - \hat{y}_{u,neg})$$

在这个过程中，Sigmoid激活函数⁺σ被用来调整输出，确保最后的输出在0和1之间，可以解释为正例和负例之间的偏好值。通过将这个成对的BPR损失整合进推荐模型的标准训练中，实现了该损失函数能够引导模型的反向传播过程，从而将计算出的梯度应用于对冷启动内容进行表示学习的嵌入参数上。

特别地，成对损失提供了额外的训练信号，对于那些在常规数据集中较少被提及或交互较少的冷启动内容格外有用。常规的推荐任务可能不足以充分训练这些内容的嵌入向量，因为它们的交互历史信息不够丰富。通过实施这个成对比较的辅助损失，即使对于这些训练不足的冷启动内容，模型也能通过成对比较的方式学习到用户的偏好模式，进而增强其泛化能力，特别是针对潜在冷启动场景的能力。

Experiments

数据与预处理。

我们使用公开的亚马逊评论数据集来评估我们方法的性能。我们选择了"*美容*"和"*运动与户外用品*"类别，这些类别包括用户对这两类物品的评分和评论。为了划分训练集和测试集⁺，我们遵循单点分割策略，并按照7:3的比例选择时间点进行数据分割。分割时间点之前的数据用于模型训练，训练后的模型在分割时间点之后的数据上进行测试。具体而言，我们将仅出现在测试数据⁺中的物品视为"*冷启动项目*"，而其他物品则视为"热启动⁺"项目，这与实际设置相似。例如，在"*运动与户外用品*"类别中，分割后得到55,255个热启动物品和2,751个冷启动物品。该类别中总共有224,956个用户查询。我们随机选取一部分用户查询，并为每个查询随机选取两个冷启动物品，生成增强数据示例。

模型与参数。 我们使用两种已建立的推荐系统后端来评估我们提出的方法的一般性：神经矩阵分解（**NeuMF**）和**SASRec** 。NeuMF从物品ID中学习用户嵌入$\mathbf{v}_u$和物品嵌入$\mathbf{v}_i$，而SASRec通过自我注意力层在用户的历史交互上编码序列信息来编码用户嵌入$\mathbf{v}_u$。这些后端代表了许多推荐系统的核心组件，我们测试了这三个基于它们的变体：

1）***无增广***基础模型，该模型使用原始训练数据进行推荐系统训练；

2）***基于内容***方法使用物品的袋模型表示，并整合其类别和标题，这是处理冷启动项目的一种常见方法；

3）***带有增广***的LLM（**LLM增广**），整合LLM生成的增广并补充训练过程中的二元比较损失。我们使用具有不同模型大小（即*XXS*、*S*和*L*）的***PaLM2***来研究合成数据生成上的性能。

K项评估指标。 对于每个测试用户查询，我们从集合$\mathcal{I} = \mathcal{I}_{cold} \cup \mathcal{I}_{warm}$中检索与最高兼容性评分匹配的前K个物品。为了离线比较推荐系统的性能，我们采用召回率⁺（即*R\@K*）来检查测试查询的地面真值物品是否出现在TOP-K列表中。我们将结果按地面真值⁺标签分组，即当购买的物品为冷启动项目时，结果在"冷启动"列中分组，其余在"热启动"列中（表 中）。

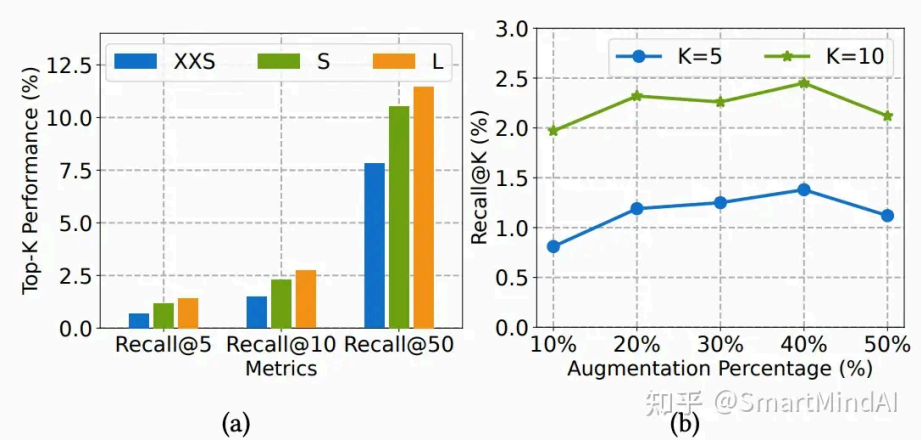
Task		Cold-start						Warm-start					
Dataset		**Beauty**			**Sports**			**Beauty**			**Sports**		
Metrics		R@5	R@10	R@50	R@5	R@10	R@50	R@5	R@10	R@50	R@5	R@10	R@50
NeuMF	w/o aug	0.14	0.22	0.48	0.01	0.02	0.13	3.44	5.36	18.76	2.65	4.76	17.98
	content-based	0.45	1.07	2.13	0.09	0.19	0.87	2.48	4.02	16.89	1.77	3.23	16.01
	w/ aug	1.19	2.32	10.53	0.22	0.41	2.11	3.35	5.21	18.03	2.32	4.75	17.67
SASRec	w/o aug	0.18	0.48	0.74	0.10	0.22	0.31	4.25	6.18	19.87	3.57	5.48	19.01
	content-based	0.56	1.26	4.77	0.15	0.30	0.96	2.90	5.43	16.89	1.89	3.10	15.43
	w/ aug	1.34	2.47	11.40	0.37	0.61	2.41	4.30	6.11	19.79	3.51	5.39	18.95

Results and Analysis

知乎

在表中，我们报告了测试在NeuMF和SASRec上的有效性的结果。在没有任何增广数据示例的情况下，冷启动项目的训练信号不足，导致所有K值*的召回极低。而基于内容的方法利用物品的标题和描述可以为冷启动项目提供不错的性能，但它忽略了协作信号，显著影响了热启动项目的推荐。相比之下，通过成对损失学习的冷启动项目的增广训练信号，可以对NeuMF和SASRec的表示学习产生益处，并在冷启动推荐上显著提升性能。此外，增广信号和成对比较损失在较高K值上提高了召回率，因为它们使模型能够排名更多的冷启动项目，包括一些不太相关但仍然符合用户喜好的项目。

结果表明，LLM是一种有效的方法来填充冷启动项目上的缺失知识。尽管冷启动项目的增广训练信号对某些召回指标对热启动项目推荐产生轻微负面影响，但与冷启动推荐的巨大收益相比，性能的下降是微不足道的。



为了解不同大型语言模型(Large Language Model, LLM)型号的影响，我们在Amazon-Beauty上采用基于NeuMF的推荐系统，使用不同型号的*PaLM2*（即XS、S和L型号）作为数据增强器。在图(a)中，我们发现模型尺寸确实影响了增强性能。众所周知，随着LLM规模的增加，它们的许多能力是涌现出来的。

我们假设较大的模型能够更准确地推断用户的历史行为和偏好。在图(b)中，我们还观察到，通过生成更多的用户查询来生成增强的训练信号，我们可以进一步提高冷启动推荐的性能。尽管增加更多的合成数据超过某一特定点（40%）后，并未带来进一步的改进。

原文《Large Language Models as Data Augmenters for Cold-Start Item Recommendation》

发布于 2024-07-10 10:57 · IP 属地北京

谷歌 (Google) LLM 冷启动