

快手破局：巧借因果链，融合搜索数据赋能推荐模型新高度



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

已关注

21 人赞同了该文章

Introduction

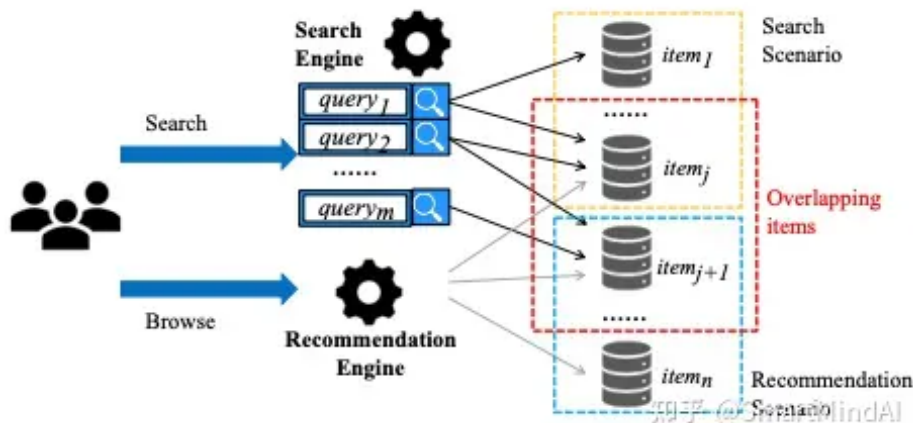
推荐和搜索是获取信息的主要方式。传统上，推荐和搜索通常作为两个分离系统，为不同的用户提供不同类型的信息目标。但现在，一些在线平台同时提供这两种服务。尽管它们有不同类型的数据输入，但它们共享一个用户和内容集。这让我们有机会通过结合两个服务收集的用户行为来提高一个服务的性能。已有研究显示联合优化搜索和推荐可以提升各因的表现。传统的推荐系统基于用户、物品和上下文信息进行推荐。这些信息通常是实数嵌入向量，因此用户的偏好根据这些嵌入来计算，比如通过用户和内容嵌入之间的点积⁺。从因果分析的角度看，嵌入信号描述的原因可以分为两部分。IVs方法可用于预测IVs影响。IV回归可将IVs分为因果相关性和非因果部分。搜索行为可以作为IVs，嵌入表示IVs并输入RS以增强用户表示。

Problem Formulation

本文将IV的推荐问题转化为搜索查询⁺的形式，并进行了形式化的处理。

Background

Recommendation and search in one platform



有许多内容平台提供搜索和推荐服务，为同一用户群体提供一组相同的产品。从推荐的角度来看，当用户 $u \in \mathcal{U}$ 访问该平台时，系统会为其提供一系列已存在的RS产品列表。通常情况下，用户 u 与某些上下文环境互动，这些上下文环境用实数向量（嵌入）

$$\mathbf{p}_u \in \mathbb{R}^d$$

示，分别表示为

$$\mathbf{t}_u \in \mathbb{R}^{d_u}$$

和

$$\mathbf{t}_i \in \mathbb{R}^{d_i}$$

其中 d_i 和 d_u 分别是用户和物品的嵌入维度。RS通常在带有历史用户-系统交互数据 \mathcal{D}^{rec} 的训练集中进行训练，其中每个元组 $(u, i, c) \in \mathcal{D}^{\text{rec}}$ 表示用户 u 被显示了物品 i ，并且交互是 $c \in \{0, 1\}$ ，其中 $c = 1$ 表示点击， $c = 0$ 否则。查询映射到一个实数向量，即

$$\mathbf{t}_q \in \mathbb{R}^{d_q}$$

其中 d_q 是嵌入维度。搜索和推荐都共享相同的用户集和物品集，因此用户和物品的嵌入向量 \mathbf{t}_u 和 \mathbf{t}_i 是相同的。历史用户-系统交互可以在

$$\mathcal{D}^{\text{src}}$$

中表示，其中每个元组 (u, q, i, c) 表示用户 u 发出查询 q 后被展示物品 i ，并且用户的行为是 $c \in \{0, 1\}$ 。

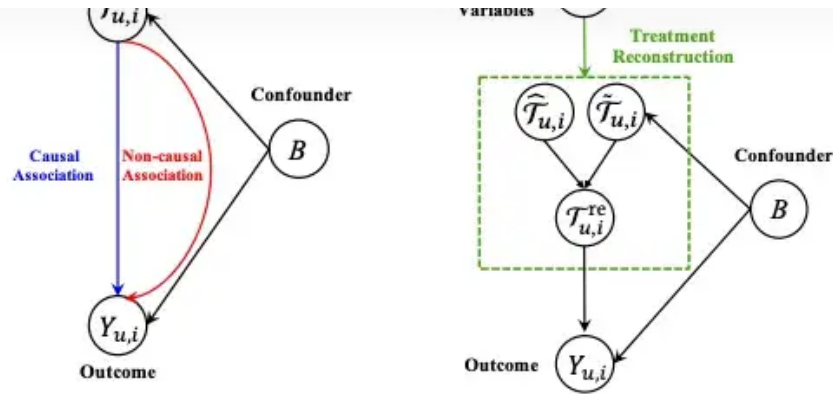
由于搜索和推荐服务使用相同的内容集来满足用户， \mathcal{D}^{rec} 和 \mathcal{D}^{src} 之间不可避免地存在重叠，即它们的记录中有共同的目标项。如图所示，在搜索和推荐场景中存在重叠的内容。

Method of instrumental variables

IV方法用于估计 因变量 X 和结果变量 Y 之间的因果效应，需要一个有效工具变量 Z 。IV方法，如2SLS，使用二阶最小二乘回归找到 因 X 对结果 Y 的影响：首先在工具变量 Z 上进行回归以获得重構的因，然后以重構后的因为第二阶段因变量进行回归，从而得到无偏的因果效应估计。

Causal view of recommendation

现有RSs通常在用户-系统历史行为中训练，其中认为在 \mathcal{D}^{rec} 中每个训练记录的点击反映了用户的偏好。但在实际应用中， \mathcal{D}^{rec} 中的用户点击往往受多种因素影响，如混淆变量、选择偏差和流行度偏差等。根据因果推断的角度，可以将用户的嵌入向量视为 因方法 $\mathcal{T}_{u,i}$ ，用户反馈（点击）视为结果 $Y_{u,i}$ 。利用给出的框架，可以构造一个关于常规RS的因果图⁺，如图1(a)所示。这个图模型⁺只是估算了 因 $\mathcal{T}_{u,i}$ 和结果 $Y_{u,i}$ 之间的混合关联。因为存在未知的混杂因素 B ，导致可能存在两种从因 $\mathcal{T}_{u,i}$ 到结果 $Y_{u,i}$ 的路径：一条是不能直接证明因果关系的非因果关系路径，这条路径由混杂因素（ $\mathcal{T}_{u,i}$ 至 $Y_{u,i}$ 的红色箭头曲线）支持；另一条是有明确因果关系路径，描述了为什么用户会偏好特定的物品（ $\mathcal{T}_{u,i}$ 至 $Y_{u,i}$ 的蓝色箭头线）。



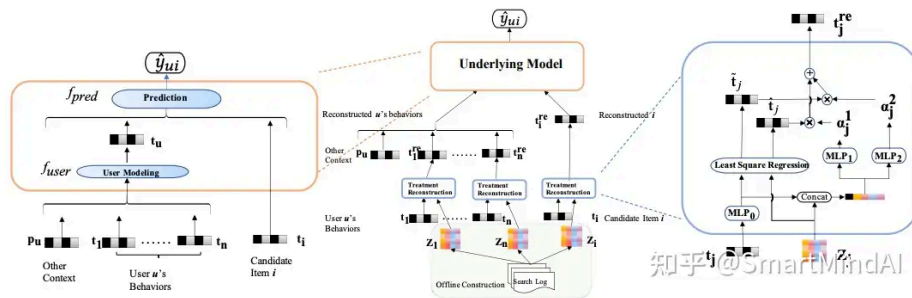
(a) Conventional recommendation models mix the causal and non-causal associations between treatment and outcome.

(b) IV4Rec reconstructs treatment by leveraging IVs to decompose treatment into causal and non-causal parts and combining them with different weights.

需要注意的是，非因果关联部分是由混杂因素影响的，例如暴露机制、公众意见、显示位置等。所以，非因果和因果关联反映了用户-内容对（即因）与用户反馈（即结果）之间的不同关系。基于偏见观察 \mathcal{D}^{rec} 难以识别因果关联。然而，用户在 \mathcal{D}^{src} 中的搜索行为提供了机会，可以通过相关查询作为IVs ($\mathcal{Z}_{u,i}$) 来分解因 $T_{u,i}$ 。然后，通过回归得到的 $\hat{T}_{u,i}$ 可以被视为与 $Y_{u,i}$ 的因果关系，且不依赖于混淆变量 B 。残差 $\tilde{T}_{u,i}$ 则反映了与 $Y_{u,i}$ 的非因果关系。因被重新组合并注入到RS中，而这个过程是在一个因果学习框架下完成的。

Our approach: IV4Rec

IV4Rec框架。



Model overview

IV4Rec包含三个步骤，如图所示。首先，它使用推荐数据 \mathcal{D}^{rec} 来定义因 $T_{u,i}$ ，并基于搜索数据 \mathcal{D}^{src} 构建IVs $\mathcal{Z}_{u,i}$ 。接下来，它通过回归方法将因 $T_{u,i}$ 与IVs $\mathcal{Z}_{u,i}$ 结合起来。最后，将重建的因输入到RS。

Construction of treatments and IVs

预测目标用户-物品对 (u, i) 的偏好得分可以通过定义RS中的一个因变量 $T_{u,i}$ ，该变量包含目标物品的嵌入和用户与之交互的物品的嵌入来实现。

$$T_{u,i} = \{t_j : j \in \mathcal{I}_u \cup \{i\}\},$$

用户 u 与物品交互时生成的嵌入向量 t_j 是 \mathbb{R}^d 上的一个元素，通常通过表示学习的方法将其投影到稠密向量中。 \mathcal{I}_u 表示用户 u 在数据集中与所有物品交互过的集合。

$$\mathcal{I}_u = \{i' : \exists (u, i', c = 1) \in \mathcal{D}^{rec}\}.$$

对于每个用户 i 和时间步数 j ,

是因组 $\mathcal{T}_{u,i}$ 的对应IVs $\mathcal{Z}_{u,i}$ 所构成的矩阵集。

$$\mathcal{Z}_{u,i} = \{\mathbf{Z}_j : j \in \mathcal{I}_u \cup \{i\}\},$$

$$\mathcal{Q}_j = \{q : \exists(u', q, j, c = 1) \in \mathcal{D}^{\text{src}}\}.$$

随后，我们根据 \mathcal{D}^{src} 中查询项 j 的点击数对 \mathcal{Q}_j 中的查询进行排序。前 N 个查询被保存为

$$\{q_k\}_{k=1}^N \subset \mathcal{Q}_j.$$

接着，为了确定物品 j 的IVs，我们可以将这 N 个查询的嵌入以堆栈⁺的方式组织起来。

$$\mathbf{Z}_j = [\mathbf{t}_{q_1}, \dots, \mathbf{t}_{q_k}, \dots, \mathbf{t}_{q_N}],$$

将向量 \mathbf{Z}_j 填充到合适大小，通过BERT模型获取查询向量 \mathbf{t}_{q_k} 。在预处理阶段，针对每项数据集项 j ，收集与之相关的查询集合 \mathcal{Q}_j 并将它们叠加到IV \mathbf{Z}_j 上以实现离线生成。

Treatment reconstruction

利用 $\mathcal{T}_{u,i}$ 和 $\mathcal{Z}_{u,i}$ 作为原始数据，我们通过回归 $\mathcal{Z}_{u,i}$ 并将其与残差相加，生成新的因 $\mathcal{T}_{u,i}^{\text{re}}$ ，图展示该过程。

Treatment decomposition

IVs目标是分离因到输出的因果关系。我们通过回归 $\mathcal{T}_{u,i}$ 与 $\mathcal{Z}_{u,i}$ 来获得无混杂变量依赖的 $\hat{\mathcal{T}}_{u,i}$ 。

$$\hat{\mathcal{T}}_{u,i} = \{\hat{\mathbf{t}}_j = f_{\text{proj}}(\mathbf{t}_j, \mathbf{Z}_j) : j \in \mathcal{I}_u \cup \{i\}\},$$

对于每个 \mathbf{t}_j 和 \mathbf{Z}_j ,

$$f_{\text{proj}}(\mathbf{t}_j, \mathbf{Z}_j) = \mathbf{Z}_j \cdot \tau_j.$$

$$f_{\text{proj}}(\mathbf{t}_j, \mathbf{Z}_j) = \mathbf{Z}_j \tau_j,$$

τ_j 是最小二乘回归的闭式解。

对于每个物品 i ，利用 \mathbf{Z}_j^\dagger 和 MLP_0 ，我们可以计算出适应向量 $\hat{\mathbf{t}}_j$ 。然后通过减去 $\hat{\mathbf{t}}_j$ ，即可得到残差部分 $\tilde{\mathcal{T}}_{u,i}$ 。

$$\tilde{\mathcal{T}}_{u,i} = \{\tilde{\mathbf{t}}_j = \text{MLP}_0(\mathbf{t}_j) - \hat{\mathbf{t}}_j : j \in \mathcal{I}_u \cup \{i\}\},$$

RS可以通过干预拟合部分和残余部分的不同方式来挖掘不同的机制，并用于预测结果。与传统IV方法不同，我们的方法使用非线性神经网络将因映射到潜在空间，具有IV方法和非线性神经网络的双重优点。

Treatment combination

重构因的公式为：

$$\hat{\mathcal{T}} = \sum_i \tilde{\mathcal{T}}_{u,i} \cdot \hat{\mathcal{T}}_{u,i}$$

$$\mathcal{T}_{u,i}^{\text{re}} = \{\mathbf{t}_j^{\text{re}} = \alpha_j^1 \hat{\mathbf{t}}_j + \alpha_j^2 \tilde{\mathbf{t}}_j : j \in \mathcal{I}_u \cup \{i\}\},$$

这两组向量对应于同一个项 j ，并且由两个MLP估计的两个组合权重分别是 α_j^1 和 α_j^2 。

$$\alpha_j^1 = \text{MLP}_1(\text{MLP}_0(\mathbf{t}_j), \mathbf{Z}_j); \quad \alpha_j^2 = \text{MLP}_2(\text{MLP}_0(\mathbf{t}_j), \mathbf{Z}_j),$$

假设我们的两个MLP输入为transformed \mathbf{t}_j 和 \mathbf{Z}_j 的拼接。在传统因果推断⁺中，我们面临的主要挑战是从观测数据中识别因果关联。因此，我们通常会丢弃残差以消除混淆变量的影响。例如，在

察结果的理解，我们可以认识到并非所有混淆变量（偏见）都应被丢弃。相反，残差可以用来改善推荐性能。

Model-agnostic application

$$\mathbf{t}_u^{\text{re}} = f_{\text{user}}(\mathbf{t}_1^{\text{re}}, \mathbf{t}_2^{\text{re}}, \dots, \mathbf{t}_n^{\text{re}}, \mathbf{p}_u),$$

$$\mathbf{t}_1^{\text{re}}, \mathbf{t}_2^{\text{re}}, \dots, \mathbf{t}_n^{\text{re}}$$

是 \mathcal{I}_u 中的重构内容向量， \mathbf{p}_u 是其他用户的表示， f_{user} 可以是任何学习用户表示的模块，如[注意力机制](#)⁺等。最后，使用学习到的用户/内容表示预测用户的偏好。

$$\hat{y}_{u,i} = f_{\text{pred}}(\mathbf{t}_u^{\text{re}}, \mathbf{t}_i^{\text{re}}),$$

f_{pred} 是一个预测偏好得分的模型，如MLP或内积。注意，训练后的因重建模块在离线模式下可以应用。这意味着，在确定了因重建模块的参数后，该模块可以用于预处理所有内容的嵌入。在线情况下，底层模型直接使用重构后的项。因此，IV4Rec在[在线推荐](#)⁺时没有额外的时间成本。

Model training

$$\mathcal{L}_{\Theta} = -\frac{1}{|\mathcal{D}^{\text{rec}}|} \sum_{(u,i,c) \in \mathcal{D}^{\text{rec}}} c \log \hat{y}_{u,i} + (1-c) \cdot \log(1 - \hat{y}_{u,i}) + \lambda \|\Theta\|^2,$$

$\hat{y}_{u,i}$ 表示第(u,i)个用户的预测偏好分数； $\|\Theta\|^2$ 代表模型的正则化项，用于防止过度拟合； $\lambda > 0$ 是一个控制[正则化](#)⁺程度的系数。

Discussion

Feasibility of using search queries as IVs

根据IVs[估计理论](#)⁺，IVs假定两个条件：外部性和相关性。对于外部性，IVs（搜索查询）与不可见的混淆变量无关。在[推荐系统](#)⁺中，常见的混淆变量包括位置偏置、选择偏置等。值得注意的是，搜索引擎和推荐服务通常部署在同一应用程序内，当用户进行搜索时发出查询，而偏差发生在用户访问RS服务时。此外，这些查询可能由访问RS的用户的其他用户发出。因此，这些搜索用户不受排名位置或曝光物品的影响。外部性意味着IVs（搜索查询）是因的原因，但不会直接影响RS的结果，即用户的点击行为。搜索和推荐有着共同的目标。

Difference with traditional IVs methods

我们引入起源嵌入作为[深度神经网络](#)⁺输入，生成神经表示嵌入。通过最小化CTR预测损失更新这一神经表示，使IVs应用成为端到端过程。此修改使模型兼具IVs优点与神经网络优点。我们在传统IVs方法中加入残差作为间接关联嵌入表示。我们关注识别和重构因中的因果与非因果部分以提高推荐精度。我们研究推荐任务中的偏见问题，使用搜索数据作为IVs，改善推荐性能并探索搜索与推荐间的因果关系。提出的模型被认为是一种使用搜索数据进行因果学习的推荐框架，能够调整观察到与未观察到的混淆效应。

Experiments

我们展示了实验结果。脚注：代码在 github.com/Ethan00Si/In...获取。

Experimental settings

Datasets

IV4Rec使用了搜索日志和推荐日志，并创建了两个数据集：一个是基于Kuaishou短视频应用程序的日志收集的；另一个是基于公开可用的MIND数据集构建的。图展示了这两个数据集中的一些统计信息。Kuaishou数据集包含12,000个用户在名为Kuaishou的应用上的行为，这些行为包括同



集按照时间顺序划分为三个子集，前5天用于训练，第6天用于验证，最后一天用于测试，最小批大小为50。利用BERT模型生成标题和摘要的组合作为查询输入，并生成768维的查询嵌入。同样地，我们使用BERT生成768维的内容嵌入。因为MIND没有带有标签的测试集⁺，所以我们在实验中使用原始的训练和验证数据作为训练和测试集，批大小为512。

Dataset	User	Item	Query	Interaction
Kuaishou	12,000	3,053,966	162,624	4,001,613
MIND	736,349	130,380	130,380	95,447,571

Baselines and evaluation metrics

IV4Rec是一种模型无关的推荐方法，适用于多种基线并提升性能。其主要原理是采用注意力多视图学习框架来聚合用户的异质行为，并通过移除不适用于某些数据集的部分模块来适应不同的数据集。IV4Rec还与JSR进行了对比，JSR则是一个通用的联合训练框架，通过优化联合损失来训练独立的搜索模型和推荐模型，但JSR有两个版本，一个是采用完全连接的前馈神经网络⁺的通用搜索组件，另一个则是基于NRHUB或DIN的推荐组件。

Implementation details

我们使用网格搜索，通过从 $\{1e-4, 3e-4, 5e-4, 7e-4, 1e-3\}$ 和 $\{0.5, 0.9, 1.0\}$ 中选择学习率和丢弃保持概率来优化神经网络的超参数。Adam用于优化。我们通过在《数据集》中的相关查询上构建IVs，将 N 设置为10来处理稀疏性问题。在MIND数据集中， N 设置为1。

Experimental results

IV4Rec框架在两组数据上均比基础模型NRHUB和DIN更有效，这支持了其在改善推荐模型上的有效性。此外，IV4Rec-NRHUB和IV4Rec-DIN也优于JSR-NRHUB和JSR-DIN，这些方法同时优化了搜索和推荐。IV4Rec-NRHUB和IV4Rec-DIN在Kuaishou数据集上的表现优于基础模型是因为它们使用搜索行为作为用户建模，而DIN则通过增加用户搜索历史特征来增强性能。这些结果表明使用搜索查询作为IV能够有效地恢复因效果。

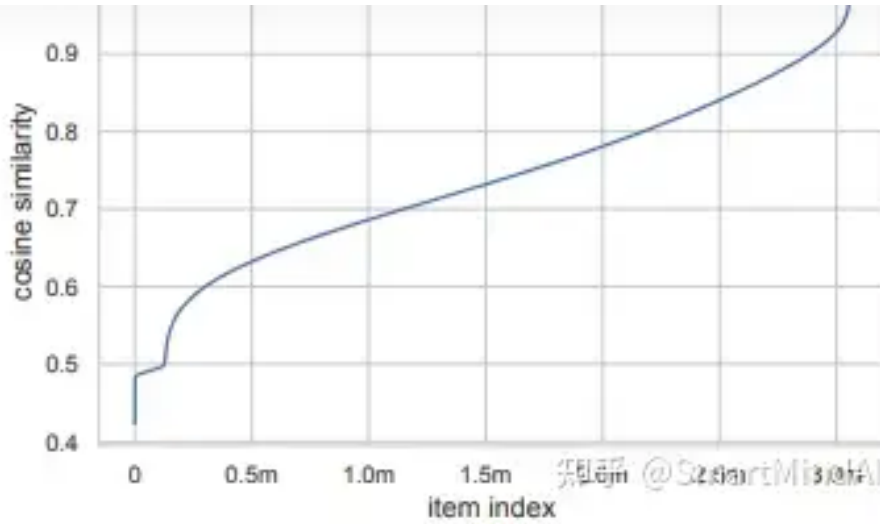
Model	Kuaishou Dataset				MIND Dataset			
	AUC	MRR	nDCG@5	nDCG@10	AUC	MRR	nDCG@5	nDCG@10
NRHUB	0.6455	0.1816	0.4347	0.4692	0.6595	0.3123	0.3428	0.4065
JSR-NRHUB	0.6488	0.1812	0.4326	0.4687	0.6660	0.3164*	0.3480*	0.4117*
IV4Rec-NRHUB	0.6574*	0.1837*	0.4411*	0.4774*	0.6722*	0.3271*	0.3609*	0.4219*
DIN	0.6512	0.1833	0.4416	0.4743	0.6851	0.3326	0.3680	0.4304
JSR-DIN	0.6524	0.1838	0.4417	0.4755	0.6873	0.3357	0.3686	0.4303
IV4Rec-DIN	0.6561 [†]	0.1844	0.4432 [†]	0.4779 [†]	0.6898 [†]	0.3336	0.3700 [†]	0.4326 [†]

Detailed empirical analysis

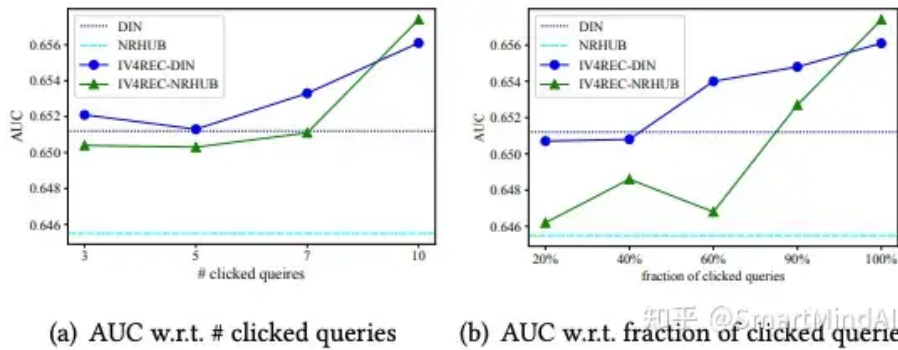
IV4Rec提升推荐准确度的详细实验证明了其有效性与原因。

Effects of search queries as IVs.

验证Section 中的相关性假设，使用Kuaishou数据集实验。

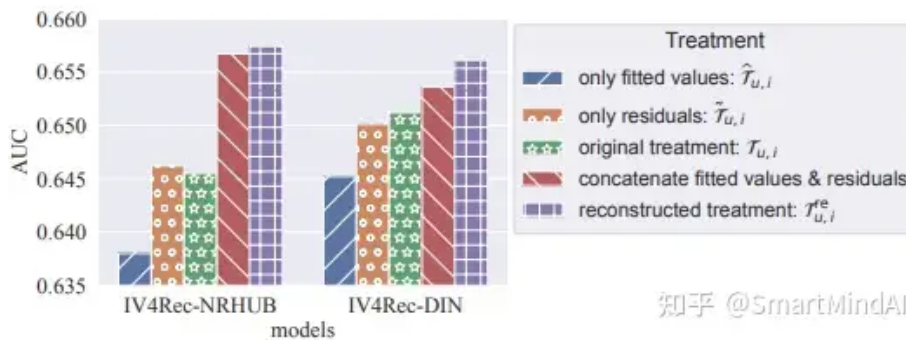


研究采用query-item对嵌入的余弦相似度*衡量项与其对应查询的相关性。嵌入由平台上的预训练模型生成，图展示了每项及其最高排名的查询的相关性。大部分相似分数大于0.6，说明大多数项与对应的查询高度相关。



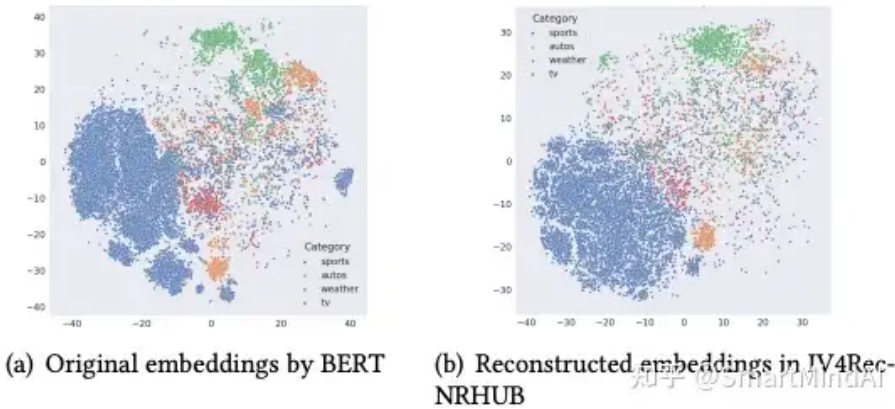
Effects of using residuals in recommendation.

本文研究了IV4Rec模型，该模型结合了拟合部分和残差部分以提高AUC。实验在Kuaishou数据集上进行，使用不同的构造方法对比：仅使用 $\hat{T}_{u,i}$ 的拟合值；仅使用 $\tilde{T}_{u,i}$ 的残差；使用原始因 $T_{u,i}$ ；使用由IV4Rec重构的 $T_{u,i}^{re}$ ；以及使用 $\hat{T}_{u,i}$ 和 $\tilde{T}_{u,i}$ 重建的因。结果显示，拟合部分和残差部分都具有推荐的贡献，且将两者结合起来作为重建的因能显著提高AUC。此外，当NRHUB和DIN作为IV4Rec的底层模型时，这一现象更为明显。研究还发现，即使非因果关联的残差部分也能够用于用户偏好预测，因为残差仍然与结果有强相关性。最后，IV4Rec提出了一种使用加权组合和两个MLPs来估计权重的方法，此方法比简单串联更优，并验证了因重建方法的有效性。



Enhancing the item embeddings.

我们在MIND数据集上进行实验，通过t-SNE可视化每篇文章的原始嵌入 t_j ，并将不同类别的文章用四种颜色表示，如图(a)所示。然后，我们利用IV4Rec-NRHUB计算这些文章的重构项嵌入 t_j^{re} ，并在图(b)中表示。通过对比这两个图，我们可以看出重构的嵌入比原始的嵌入更加分散且聚类效果更好。例如，“体育”类别的文章在图(b)中的分布更为紧密，集中在底部左侧。因此，结果表



Conclusions

提出了一种名为IV4Rec的模型无关的因果学习框架，使用搜索数据来改善推荐。该框架利用搜索查询作为IV，将其推荐分为因果关联部分和非因果关联部分，以便了解其不同机制以进行偏好预测。同时，IV4Rec结合了传统方法（工具变量）与深度神经网络，提供了端到端的框架来估计模型参数。在Kuaishou产品数据和公开基准上的实验证明了IV4Rec的有效性。

论文原文《A Model-Agnostic Causal Learning Framework for Recommendation using Search Data》

发布于 2024-01-23 14:01 · IP 属地北京

快手 因果推理 搜索推荐系统

赞同 21 添加评论 分享 喜欢 收藏 申请转载 ...

理性发言，友善互动



还没有评论，发表第一个评论吧

推荐阅读



快手 | DataFunSummit

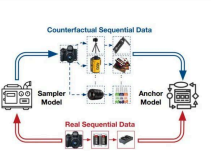
Causal Machine Learning in User Growth

李强 博士 快手 数据科学家 & 算法Leader



美团

IDENTIFY CAUSAL EFFECT



Counterfactual Sequential Data

Sampler Model

Anchor Model

Real Sequential Data