# 谷歌STAR:一种无需训练的LLM4Rec方法

原创 州懂学习笔记 州懂学习笔记 2024年12月14日 09:41 广东

州懂学习笔记
分享大模型推荐系统相关知识和学习笔记
53篇原创内容

公众号

## 谷歌STAR:一种无需训练的LLM4Rec方法

> 标题: STAR: A Simple Training-free Approach for Recommendations using Large Language Models
> 地址: https://arxiv.org/pdf/2410.16458
> 公司: Google

## 1. 前言

随着近年来LLM的突破性进展, 业界也在不断尝试将LLM应用于推荐系统。当前, LLM4Rec的范式主要可以分为以下几类:

### 1) LLM as a Feature Encoder

LLM有着非常强大的内容理解能力, 将LLM看作是特征Encoder可以从Item的元数据以及用户画像中捕获丰富的语义信息, 这类方法主要包括:

1. 将连续的LLM Embed离散化(如vector量化)处理，并训练后续的生成模型;
2. 使用LLM Embed初始化序列模型的Embed参数;
3. 训练模型直接计算Item和用户Embed间的相关性。

虽然优化这些Embed可以提高推荐整体性能，但这通常需要增加训练成本, 且牺牲了一定的通用性。

### 2) LLM as a Scoring and Ranking function

最近的一些研究表明，LLM可以通过基于自然语言(通过生成Prompt)的方式理解用户偏好或过去的交互来推荐Item。然而，这些研究表明，单独使用LLM不如利用协同知识的User-Item

交互数据微调的模型有效。为了弥合使用协同知识和使用LLM语义理解之间的差距，近期业界的重点是在利用交互数据微调模型上，尽管这种方法的成本也很高。

### 3) LLM as a Ranker for Information Retrieval

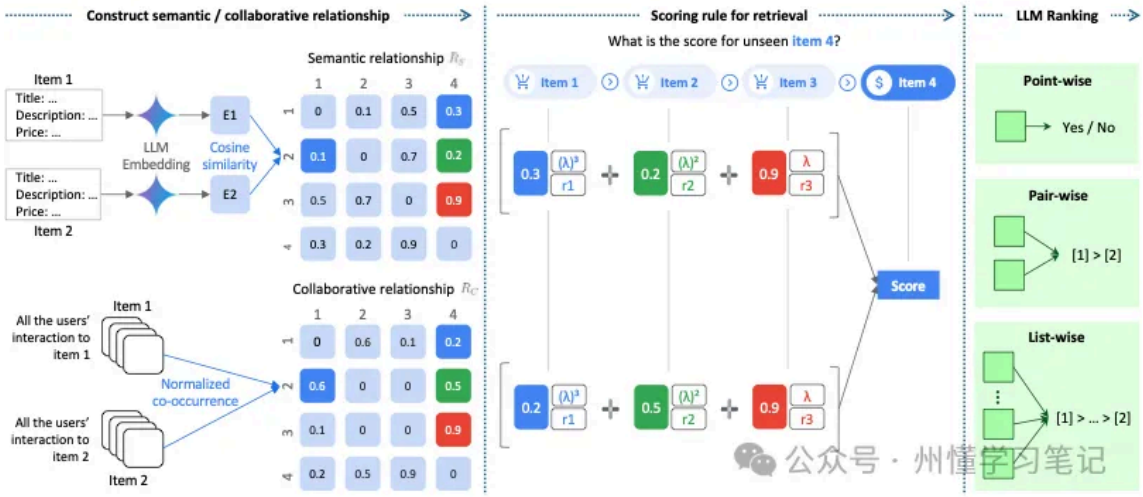
这种将LLM直接作为Ranker的方式常用于文档检索中, 最近的一些研究表明, LLM在zero-shot的效果甚至超过了传统的有监督cross-encoder方式。这类方法的提示可以分成3类:

- point-wise：LLM直接用于评估相关性，但这种方法很难捕捉到文档的相对重要性
- pair-wise：LLM直接对比Item Pair对之间的相对偏好, 这是更有效的方法, 但效率低下, 因为需要涉及大量调用
- list-wise：LLM同时对比多个Item，但性能很大程度上严重依赖于模型的语义先验和推理能力。

本文作者主要研究LLM在推荐场景下的排序能力, 这些任务本质上是主观的, 这与文档检索的确定性性质不同, 此外, 大多数方法都需要使用下游任务数据微调LLM, 会带来一些额外的训练成本。本文作者提出了一种简单&无需训练的框架(Simple Training-free Approach for Recommendation，STAR), 包括检索和排序两部分, 下面详细介绍。

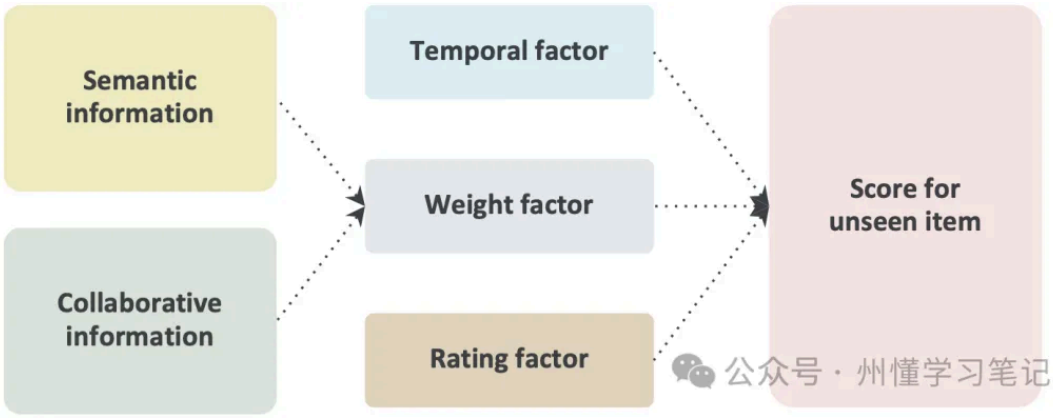
## 2. 方法

STAR模型的整体框架如下图所示, 包括检索和排序两部分:




### 2.1 检索部分

检索部分是基于用户$u$的历史行为序列$S_u = \{s_1, s_2, \ldots, s_n\}$给未见过的Item$x \in I$打分。该部分的整体思路如下图所示, 通过打分规则, 将语义信息和协同信息融合进权重因子中, 再与时序因子和评分因子相结合去使用, 无需额外的微调训练。




### 2.1.1 语义关系

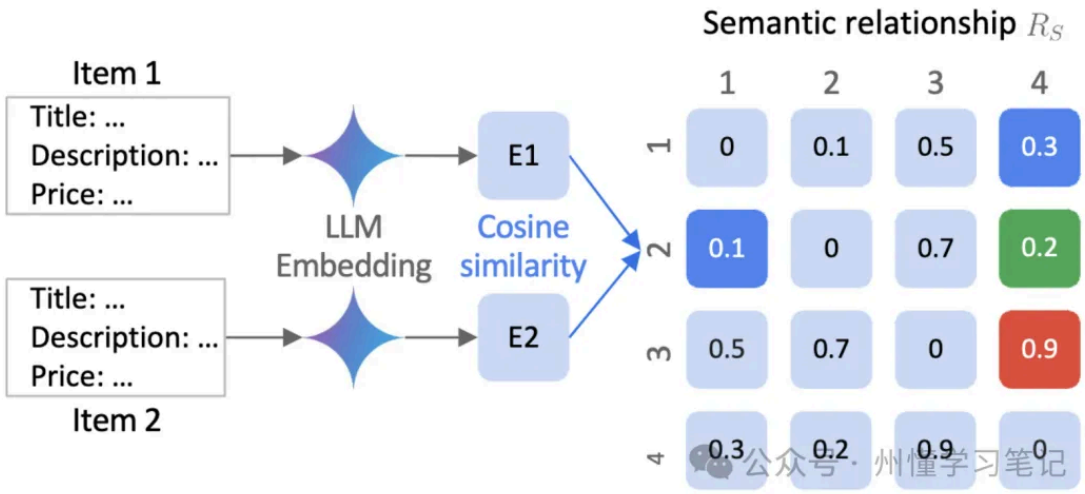
作者首先将Item的标题、描述、类目、品牌、销售量排名以及价格等信息都输入到提示中, 再通过LLM的Embedding API获取语义向量。下图给出了一种提示的示例:

```
description:
    LENGTH: 70cm / 27.55 inches
    Color: Mix Color
    EST. SHIPPING WT.: 310g
    Material: Synthetic High Temp Fiber
    Cap Construction: Capless
    Cap Size: Average
    1. The size is adjustable and no pins or tape should be required. It should fit most people.
       Adjust the hooks inside the cap to suit your head.
    2. Please be aware that colors might look slightly different in person due to camera quality and monitor settings.
       Stock photos are taken in natural light with no flash.
    3. Please ask all questions prior to purchasing. I will replace defective items.
       Indicate the problem before returning. A 30-day return/exchange policy is provided as a satisfaction guarantee.

title: 63cm Long Zipper Beige+Pink Wavy Cosplay Hair Wig Rw157
salesRank: {'Beauty': 2236}
categories: Beauty > Hair Care > Styling Products > Hair Extensions & Wigs > Wigs
price: 11.83
brand: Generic
```



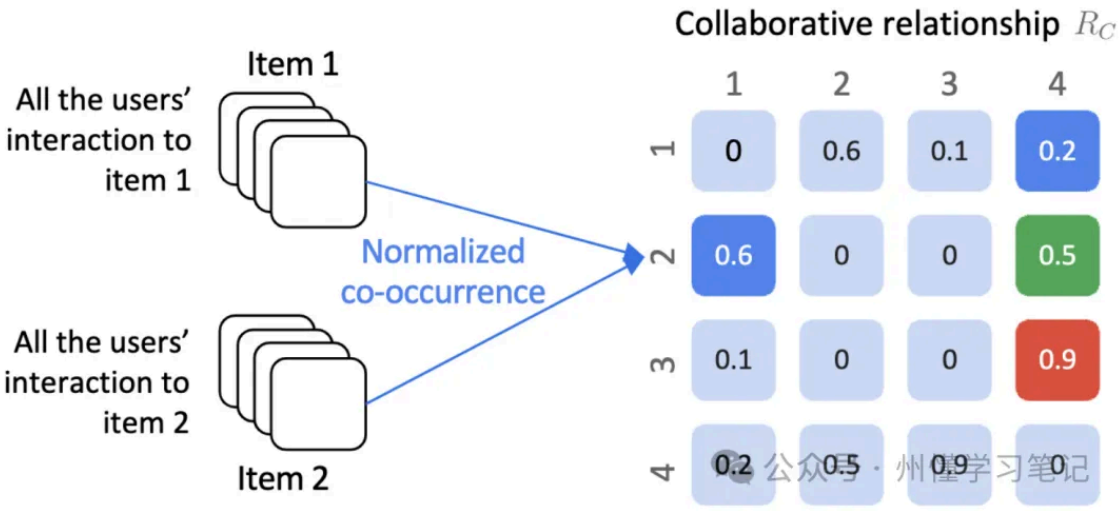
通过这种方式可以提前计算全量Item集的Embedding, 然后就可以使用余弦相似度计算用户行为序列中的Item(如下图的#1~3号)与候选Item(如下图#4号)的语义相似度得分, 记为$R_C$:




### 2.1.2 协同关系

协同共现信息是推荐系统的核心, 这里, 作者提前计算构建共现得分矩阵。具体地, 作者统计出每个Item的用户交互, 得到User-Item交互矩阵$C \in \mathbb{R}^{n \times m}$(如0-1值表示该User和Item是否有

过交互), 其中$n$是Item数,$m$是用户数, 再通过余弦相似度计算两两Item的得分得到协同共现得分, 记为$R_{\mathrm{C}}$, 如下图所示:



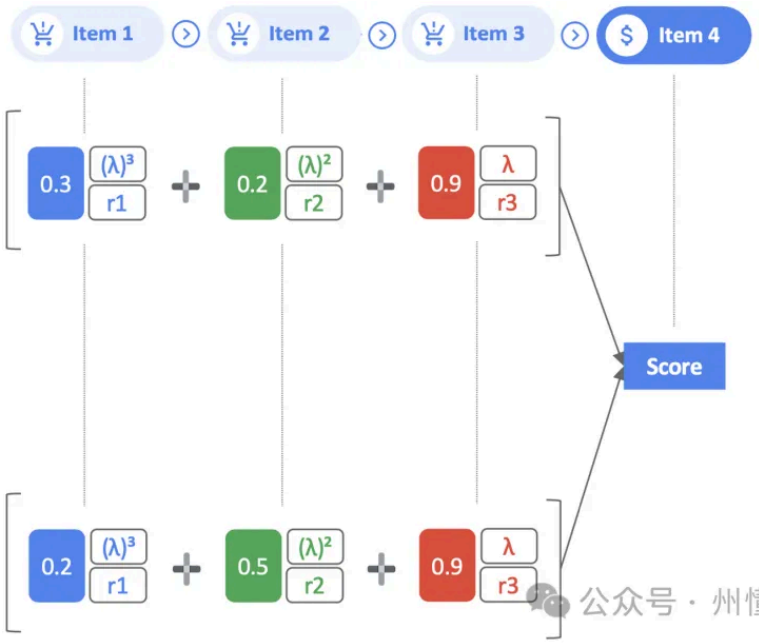

### 2.1.3 打分融合规则

在得到候选Item$x$与用户行为序列中每个Item$s_j$的语义得分$R_{\mathrm{S}}^{xj}$和协同共现得分$R_{\mathrm{C}}^{xj}$之后, 作者将这两个分数按下面的方式进行融合:

$$\texttt{score}(x) = \frac{1}{n}\sum_{j=1}^{n} r_j \cdot \lambda^{t_j} \cdot \left[a \cdot R_{\mathrm{S}}^{xj} + (1-a) \cdot R_{\mathrm{C}}^{xj}\right]$$

其中,$r_j$是用户对历史交互Item的评价得分, 比如用户对看过的电影进行评分,$\lambda^{t_j}$是以用户行为序列中的顺序为指数的时间衰减因子,$a$是融合的超参。

检索流程以得分最高的top$k$个item作为最终输出:




## 2.2 排序部分

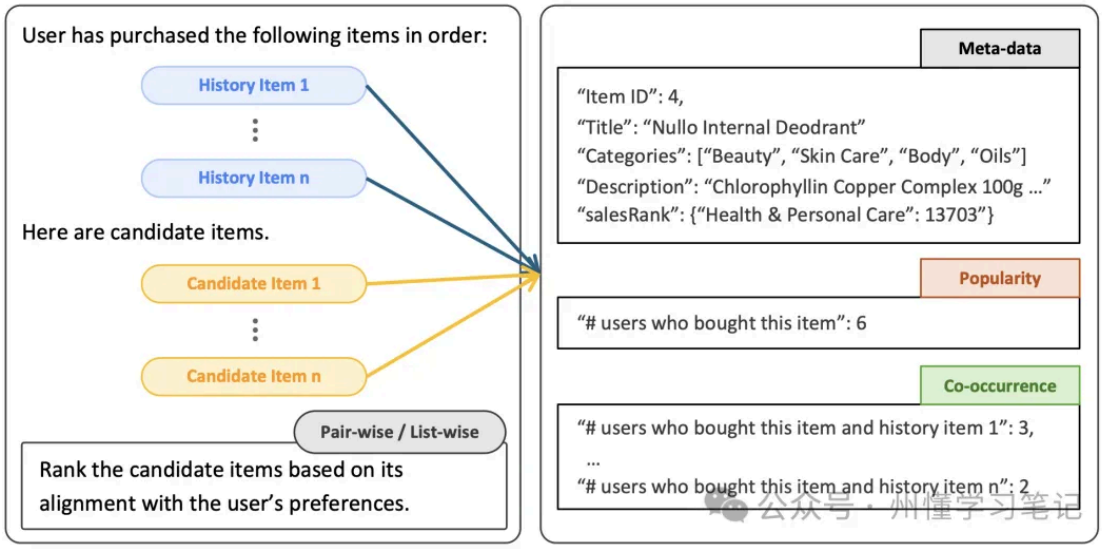
### 2.2.1 排序策略

作者将检索部分得到的Item按得分顺序作为输入, 然后使用了三种不同的排序策略:

1. **point-wise:** 基于用户序列$S_u$独立评估每个Item$x \in I_k$的得分，以确定用户$u$将与Item$x$交互的可能性有多大。如果两个Item得分相同, 则将检索的得分打的排前面。
2. **pair-wise:** 基于用户序列$S_u$评估两个Item之间的偏好。具体地, 作者使用滑动窗口方法, 对列表按检索分数从低到高比较列表, 进行位置交换。
3. **list-wise:** 也是使用滑动窗口方法, 对窗口大小$w$个Item进行比较, 并按步幅$d$进行滑动。pair-wise可以看作是($w = 2, d = 1$)的list-wise。

### 2.2.2 Item提示信息

下图给出了排序过程的prompt构造, 除了一些Item的meta信息之外, 还额外补充了流行度信息以及共现信息




1. **流行度信息:** 统计Item在数据集中的交互次数, 如"购买该Item的用户数:###"
2. **共现信息:** 统计两个Item的共现交互次数, 如"同时购买该商品和用户历史商品#1的用户数量:###"

## 3. 实验部分

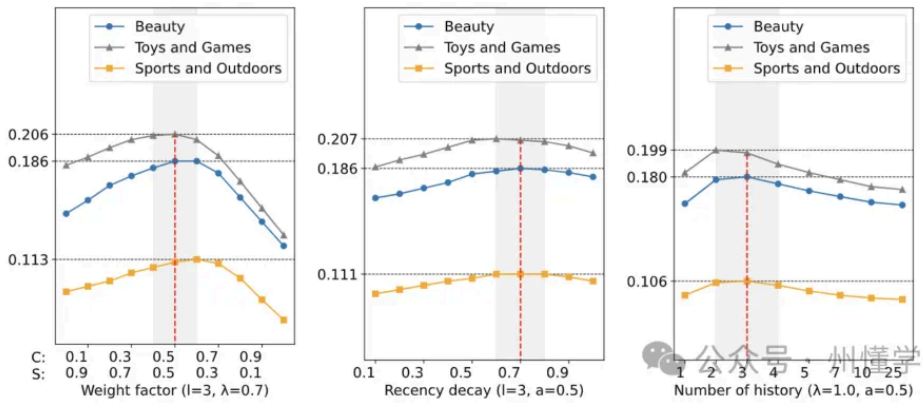
### 3.1 整体效果

与其它模型的整体效果对比

| Category | Method / Model | Train | Beauty | | | | Toys and Games | | | | Sports and Outdoors | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | H@5 | N@5 | H@10 | N@10 | H@5 | N@5 | H@10 | N@10 | H@5 | N@5 | H@10 | N@10 |
| Baseline | KNN | ✓ | 0.004 | 0.003 | 0.007 | 0.004 | 0.004 | 0.003 | 0.007 | 0.004 | 0.001 | 0.001 | 0.002 | 0.001 |
| | Caser [54] | ✓ | 0.021 | 0.013 | 0.035 | 0.018 | 0.017 | 0.011 | 0.027 | 0.014 | 0.012 | 0.007 | 0.019 | 0.010 |
| | HGN [55] | ✓ | 0.033 | 0.021 | 0.051 | 0.027 | 0.032 | 0.022 | 0.050 | 0.028 | 0.019 | 0.012 | 0.031 | 0.016 |
| | GRU4Rec [56] | ✓ | 0.016 | 0.010 | 0.028 | 0.014 | 0.010 | 0.006 | 0.018 | 0.008 | 0.013 | 0.009 | 0.020 | 0.011 |
| | BERT4Rec [9] | ✓ | 0.020 | 0.012 | 0.035 | 0.017 | 0.012 | 0.007 | 0.020 | 0.010 | 0.012 | 0.008 | 0.019 | 0.010 |
| | FDSA [57] | ✓ | 0.027 | 0.016 | 0.041 | 0.021 | 0.023 | 0.014 | 0.038 | 0.019 | 0.018 | 0.012 | 0.029 | 0.016 |
| | SASRec [58] | ✓ | 0.039 | 0.025 | 0.061 | 0.032 | 0.046 | 0.031 | 0.068 | 0.037 | 0.023 | 0.015 | 0.035 | 0.019 |
| | $S^3$-Rec [59] | ✓ | 0.039 | 0.024 | 0.065 | 0.033 | 0.044 | 0.029 | 0.070 | 0.038 | 0.025 | 0.016 | 0.039 | 0.020 |
| | P5 [59] | ✓ | 0.016 | 0.011 | 0.025 | 0.014 | 0.007 | 0.005 | 0.012 | 0.007 | 0.006 | 0.004 | 0.010 | 0.005 |
| | TIGER [40] | ✓ | 0.045 | 0.032 | 0.065 | 0.038 | 0.052 | 0.037 | 0.071 | 0.043 | 0.026 | 0.018 | 0.040 | 0.023 |
| | IDGenRec [44] | ✓ | 0.062 | 0.049 | 0.081 | 0.054 | 0.066 | 0.048 | 0.087 | 0.055 | **0.043** | **0.033** | **0.057** | **0.037** |
| STAR-Retrieval | - | ✗ | 0.068 | 0.048 | 0.098 | 0.057 | 0.086 | 0.061 | 0.118 | 0.071 | 0.038 | 0.026 | 0.054 | 0.031 |
| STAR-Ranking | point-wise | ✗ | 0.068 | 0.047 | 0.096 | 0.056 | 0.086 | 0.061 | 0.117 | 0.071 | 0.037 | 0.026 | 0.054 | 0.031 |
| | pair-wise | ✗ | **0.072** | **0.051** | **0.101** | **0.060** | **0.090** | **0.064** | **0.120** | **0.073** | 0.040 | 0.028 | 0.056 | 0.034 |
| | list-wise | ✗ | 0.065 | 0.047 | 0.090 | 0.055 | 0.083 | 0.060 | 0.111 | 0.069 | 0.036 | 0.026 | 0.052 | 0.031 |

## 3.2 超参分析

检索阶段不同超参的分析




排序策略中不同窗口大小和不同步幅的效果对比

| Prompt Style | Window Size | Stride | Beauty | | | | Toys and Games | | | | Sports and Outdoors | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | H@5 | N@5 | H@10 | N@10 | H@5 | N@5 | H@10 | N@10 | H@5 | N@5 | H@10 | N@10 |
| None (STAR-Retrieval) | - | - | 0.0684 | 0.0480 | 0.0977 | 0.0574 | 0.0857 | 0.0606 | 0.1176 | 0.0709 | 0.0379 | 0.0262 | 0.0542 | 0.0314 |
| Selection | - | - | 0.0691 | 0.0484 | 0.0958 | 0.0570 | 0.0841 | 0.0613 | 0.1109 | 0.0699 | 0.0376 | 0.0269 | 0.0520 | 0.0316 |
| Point-wise | 1 | 1 | 0.0685 | 0.0472 | 0.0956 | 0.0558 | 0.0855 | 0.0611 | 0.1170 | 0.0713 | 0.0370 | 0.0257 | 0.0539 | 0.0312 |
| Pair-wise | 2 | 1 | 0.0716 | **0.0506** | **0.1008** | **0.0600** | **0.0899** | **0.0639** | 0.1196 | **0.0734** | 0.0401 | **0.0283** | **0.0564** | **0.0335** |
| List-wise | 4 | 2 | **0.0724** | 0.0502 | 0.1002 | 0.0592 | 0.0894 | 0.0634 | 0.1195 | 0.0732 | **0.0406** | 0.0282 | 0.0559 | 0.0331 |
| | 8 | 4 | 0.0688 | 0.0484 | 0.0988 | 0.0581 | 0.0874 | 0.0625 | **0.1202** | 0.0731 | 0.0388 | 0.0276 | 0.0556 | 0.0330 |
| | 10 | 5 | 0.0676 | 0.0480 | 0.0981 | 0.0578 | 0.0853 | 0.0616 | 0.1201 | 0.0728 | 0.0379 | 0.0270 | 0.0558 | 0.0327 |
| | 20 | - | 0.0653 | 0.0471 | 0.0903 | 0.0551 | 0.0829 | 0.0603 | 0.1113 | 0.0694 | 0.0364 | 0.0262 | 0.0518 | 0.0311 |

**Table 5: Ranking performance (Hits@K, NDCG@K) by window size and stride. Here we use 20 candidates from the retrieval stage. The best prompt for each dataset is shown in bold, and the second best is underlined.**

## 3.3 消融实验

检索阶段是否利用评分信息的消融实验

| Rating | Beauty | | Toys and Games | | Sports and Outdoors | |
|---|---|---|---|---|---|---|
| | H@10 | N@10 | H@10 | N@10 | H@10 | N@10 |
| w/ rating | 0.095 | 0.056 | 0.115 | 0.069 | 0.052 | 0.030 |
| w/o rating | **0.098** | **0.057** | **0.118** | **0.071** | **0.054** | **0.031** |

关于Item提示信息的消融实验

| Item prompt | Beauty | | Toys and Games | | Sports and Outdoors | |
|---|---|---|---|---|---|---|
| | H@10 | N@10 | H@10 | N@10 | H@10 | N@10 |
| Metadata | 0.1000 | 0.0567 | 0.1193 | 0.0690 | 0.0544 | 0.0315 |
| + popularity | 0.0998 | 0.0564 | 0.1174 | 0.0701 | 0.0549 | 0.0316 |
| + co-occurrence | **0.1008** | **0.0600** | 0.1196 | 0.0734 | **0.0564** | **0.0335** |
| + popularity, co-occurrence | 0.0999 | 0.0599 | **0.1203** | **0.0736** | 0.0550 | 0.0322 |