第八部分：涨指标的方法

1、推荐系统的评价指标

2、涨指标的方法

1、推荐系统的评价指标

日活用户数（DAU）和留存是最核心的指标

目前工业界最常用LT7和LT30衡量留存

例如：某用户今天（t_0）登录APP，未来七天（$t_0 - t_6$）中有4天登录APP，那么该用户今天（t_0）的LT7等于4

其中有：1≤LT7≤7，1≤LT30≤30

LT（全体用户LT的平均）增长通常意味着用户体验提升（除非LT增长而DAU下降，好比假设APP禁止低活用户登录）

时长增长：LT通常会增长，但阅读数、曝光数可能会下降

其他核心指标：用户使用时长、总阅读数（即总点击数）、总曝光数等，这些指标重要性低于DAU和留存

非核心指标：点击率、交互率等等

（对于UGC平台，发布量和发布渗透率也是核心指标）

2、涨指标的方法

（1）改进召回模型，添加新的召回模型

（2）改进排序（粗排和精排）模型

（3）提升召回、粗排、精排的多样性

（4）特殊对待新用户，低活用户等特殊人群

（5）利用关注、转发、评论这三种交互行为

2.1 召回模型的改进

推荐系统有几十条召回通道，它们的召回总量是固定的。总量越大，指标越好，粗排计算量越大

双塔模型和item-to-item是最重要两类召回模型，占据召回的大部分配额

有些小众模型占据配额很少，在召回总量不变的前提下，添加某些召回模型可以提升核心指标

有很多内容池，如30天物品、1天物品、6小时物品、新用户优质内容池、分人群内容池等，同一个模型可以用于多个内容池，得到多条召回通道（共用一个双塔模型，因此只训练一个双塔模型）

2.1.1 双塔模型

方向1：优化正样本、负样本

简单正样本：有点击的（用户，物品）二元组

简单负样本：随机组合的（用户，物品）二元组

困难负样本：排序靠后的（用户，物品）二元组

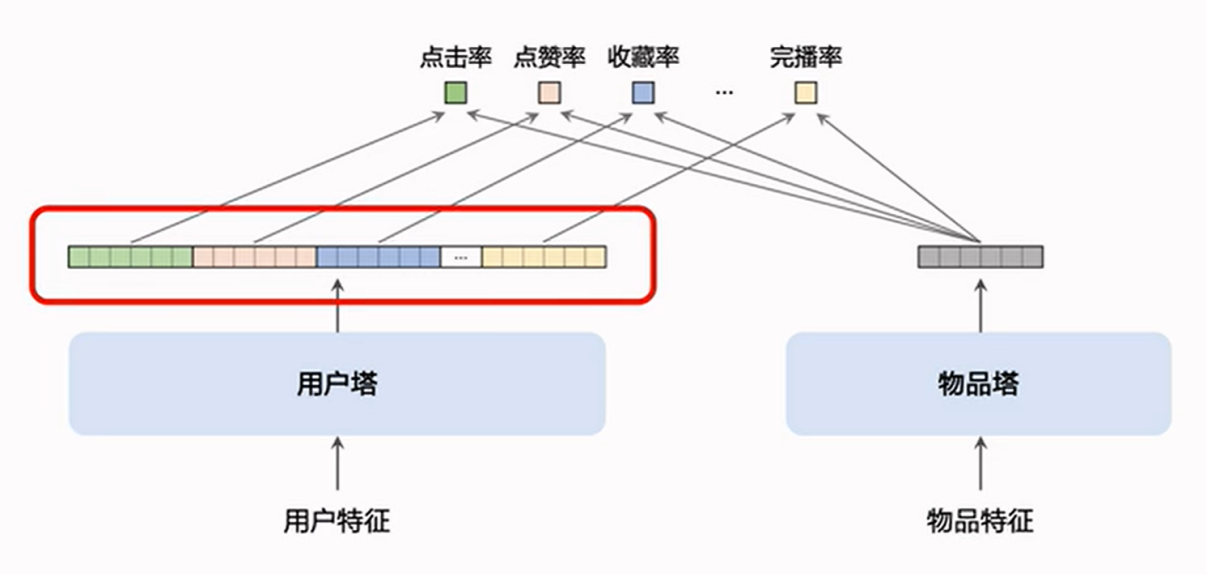
方向2：改进神经网络结构

Baseline：用户塔、物品塔分别是全连接网络，各输出一个向量，分别作为用户、物品的表征

改进1：用户塔、物品塔分别用DCN代替全连接网络（第四章内容）

改进2：在用户塔中使用用户行为序列（last-n，第五章内容）

改进3：使用多向量模型代替单向量模型（标准双塔模型是单向量模型）



如图，用户塔输出很多向量（类似多目标排序模型），分别表示点击率、点赞率、收藏率、完播率等指标

方向3：改进模型的训练方法

Baseline：做二分类，让模型学会区分正样本和负样本

改进1：结合二分类、batch内负采样（需要纠偏）

改进2：采用自监督学习方法，让冷门物品embedding学得更好

2.1.2 Item-to-Item（I2I）

I2I：基于相似物品做召回的一大类模型

最常见用法：U2I2I（user→item→item）

过程：用户u喜欢物品 i_1（用户历史交互过的物品），于是寻找 i_1的相似物品 i_2（即I2I），将 i_2推荐给u

如何计算物品相似度?

方法1：ItemCF及其变体（如ItemCF、Online ItemCF、Swing、Online Swing都是基于相同的思想），线上同时使用上述4种I2I模型，各分配一定配额

方法2：基于物品向量表征，计算向量相似度（双塔模型、图神经网络均可计算物品向量表征）

2.1.3 小众召回模型

U2U2I（user→user→item）：已知用户 u_1与 u_2相似，且 u_2喜欢物品i，那么给用户 u_1推荐物品i

U2A2I（user→author→item）：已知用户u喜欢作者a，且a发布物品i，那么给用户u推荐物品i

U2A2A2I（user→author→author→item）：已知用户u喜欢作者 a_1，且作者 a_1和作者 a_2相似，作者 a_2发布过物品i，那么给用户u推荐物品i

更复杂的模型

Path-based Deep Network (PDN)

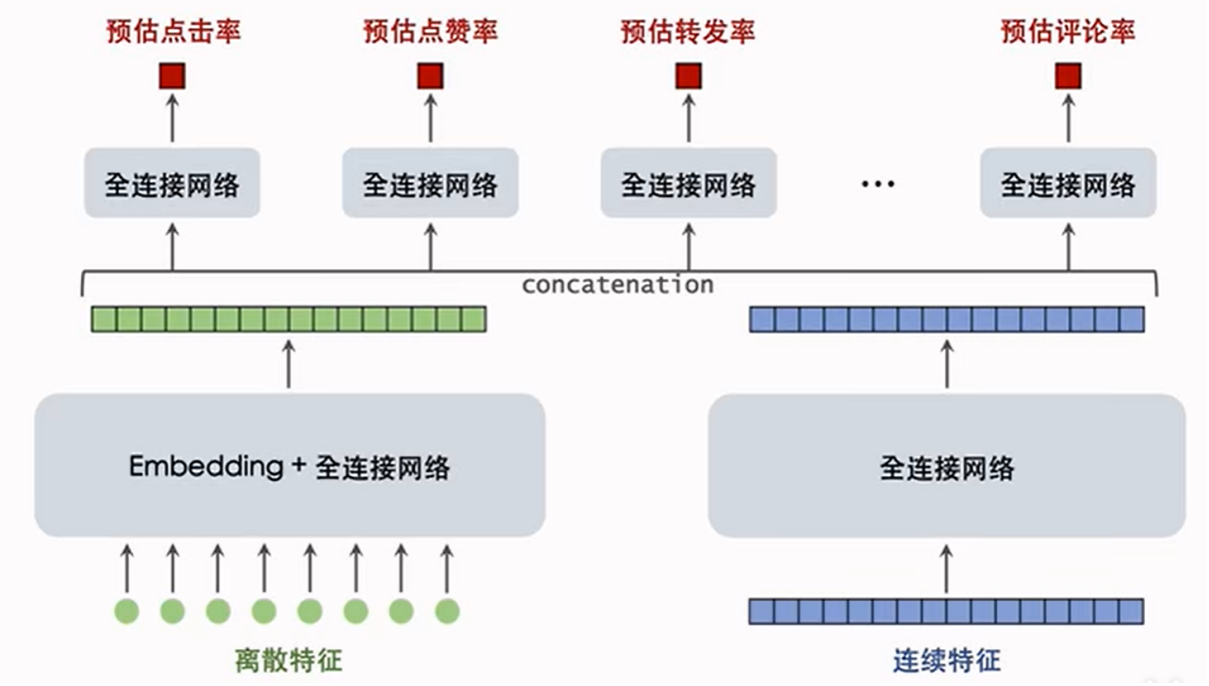
Deep Retrieval

Sparse-Interest Network (SINE)

Multi-task Multi-view Graph Representation Learning (M2GRL)

2.2 排序模型的改进

2.2.1 精排模型的改进



下面两个神经网络为基座，把离散/连续特征映射到数值向量

基座和上层的多目标预估部分都有很多优化点

基座

基座的输入包括离散特征和连续特征，输出一个向量，作为多目标预估的输入

改进1：基座加宽加深（针对全连接网络，神经网络99%的参数是在embedding层），计算量更大，预测更准确

改进2：做自动的特征交叉，如bilinear和LHUC

改进3：特征工程，比如添加统计特征、多模态内容特征

多目标预估部分

改进1：增加新的预估目标，并将预估结果加入融分公式

最标准的目标包括点击率、点赞率、收藏率、转发率、评论率、关注率、完播率.....

寻找更多新的目标，比如进入评论区、给他人写的评论点赞.....

改进2：MMoE、PLE等结构可能有效，但往往无效

改进3：纠正position bias可能有效，但可能也无效

2.2.2 粗排模型的改进

粗排打分量比精排大10倍，因此粗排模型必须够快

简单模型：多向量双塔模型，同时预估点击率等多个目标

复杂模型：三塔模型效果好，但工程实现难度较大

粗精排一致性建模

蒸馏精排模型训练粗排，让粗排和精排更一致，这样可以提升核心指标

方法1：pointwise蒸馏

设y是用户真实行为，设p是精排的预估，用（y+p）/2作为粗排拟合的目标

方法2：pairwise或listwise蒸馏

给定k个候选物品，按照精排预估做排序

做learning to rank（LTR），让粗排拟合物品的排序（而非值）

例如：对物品i和j，精排预估点击率为 $p_i > p_j$，那么LTR鼓励粗排预估点击率满足 $q_i > q_j$，否则给予惩罚

LTR通常使用pairwise logistic loss

粗精排一致性缺点：如果精排出bug，精排预估值p有偏，会污染粗排训练数据

2.2.3 用户行为序列建模

方法：简单平均、DIN、SIM

改进1：增加序列长度，让更多的交互笔记反映用户行为特点，可以使预测更准确，但计算成本和推理时间增加

改进2：筛选的方法，比如用类目、物品向量表征聚类

改进2具体做法：

离线用多模态神经网络提取出物品内容特征，将内容表征为向量；

离线将物品向量聚为1000类，每个物品有一个聚类序号；

例如：线上排序时，用户行为序列中有n=1000000个物品，某候选物品的聚类序号是70，对n个物品做筛选，只保留聚类序号为70的物品，n个物品中只有数千个被保留下来

同时有好几种筛选方法，取筛选结果的并集

改进3：对用户行为序列中的物品，使用ID以外的一些特征

工业界的做法概括：沿着SIM的方向发展，让原始的序列尽量长，然后做筛选降低序列长度，最后将筛选结果输入DIN

2.2.4 在线学习

模型更新（参考第二章）：全量更新和增量更新，增量更新即在线学习

在线学习的资源消耗

既需要在凌晨做全量更新，也需要全天不间断做增量更新，因此需要额外算力

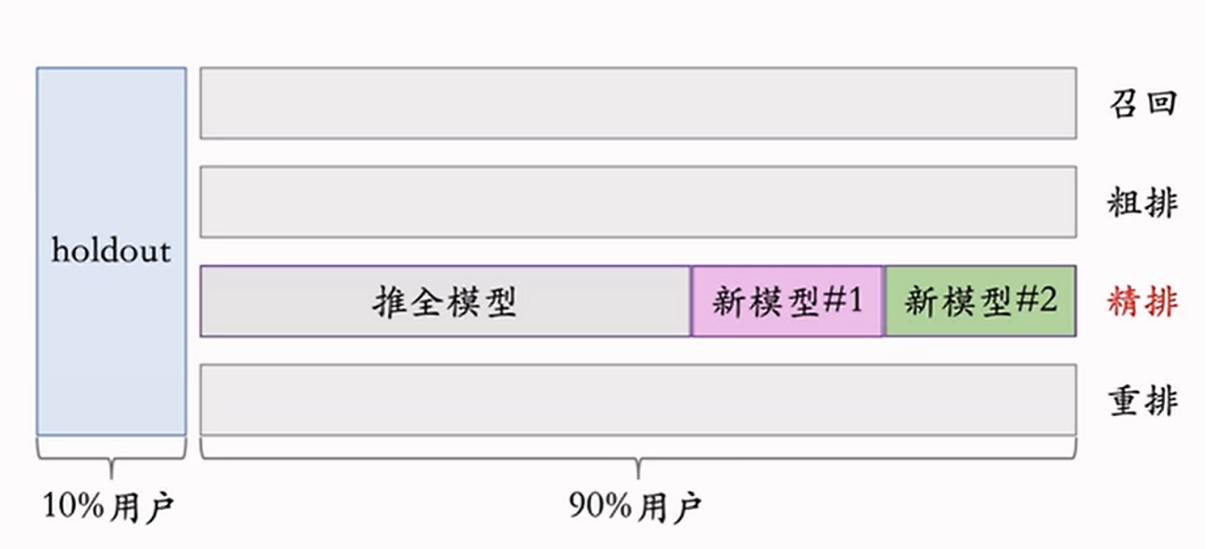
推荐系统一共需要多少额外算力给在线学习？

为了做AB测试，线上同时运行多个不同的模型：

如果线上有m个模型，则需要m套在线学习的机器

线上有m个模型，其中1个是holdout，1个是推全的模型，m-2个测试的新模型

（每套在线学习的机器成本都很大，因此m数量很小）



总结：在线学习对指标提升很大，但是会降低模型迭代升级效率

2.2.5 老汤模型

内容：

用每天新产生的数据对模型做1 epoch 的训练，久而久之，老模型训练的非常好，很难被超过；而对模型做改进重新训练，很难追上老模型

问题1：如何快速判断新模型结构是否优于老模型？

（不需要追上线上的老模型，只需要判断新老模型谁的结构更优）

对于新模型结构：

全连接层：和老模型一样都随机初始化

embedding层：可以随机初始化，也可以复用老模型训练好的参数

用n天的数据训练新老模型（从旧到新，训练1 epoch），如果新模型显著优于老模型，新模型很可能更优

问题2：如何更快追平、超过线上的老模型？

（由问题1得到初步结论：新模型很可能优于老模型，那么我们需要实现利用几十天的数据，新模型就能追上训练上百天的老模型）

方法1：尽可能多的复用老模型训练好的embedding层，避免随机初始化（embedding层是对物品、用户特点的记忆，比全连接层学得慢）

方法2：用老模型做teacher，蒸馏新模型（用户真实行为是y，老模型预测是p，用（y+p）/ 2作为训练新模型的目标）

2.3 提升多样性

2.3.1 精排多样性

（1）精排阶段，**结合兴趣分数和多样性分数**对物品i排序：

s_i：兴趣分数，即融合点击率等多个预估目标

d_i：多样性分数，即物品i与已经选中的物品的差异

用$s_i + d_i$对物品做排序

计算多样性分数常用方法：MMR、DPP等

精排使用滑动窗口，粗排不使用滑动窗口

原因：精排决定最终的曝光，曝光页面上邻近物品相似度应该小，因此要确定一个窗口内的多样性要好；而粗排考虑的是整体多样性，而非一个滑动窗口中的多样性

（2）除了多样性分数，精排还**使用打散策略**增加多样性

类目：当前选中物品i，之后5个位置不允许跟i的二级类目相同

多模态：事先计算物品多模态内容向量表征，将全库物品聚为1000类；在精排阶段，如果当前选中物品i，之后10个位置不允许跟i同属一个聚类

2.3.2 粗排多样性

提升粗排和精排多样性都可以提升推荐系统核心指标

步骤：

（1）粗排给5000个物品打分，选出500个物品进入精排

（2）根据 s_i对5000个物品排序，分数最高的200个物品送入精排；

（3）对于剩余的4800个物品，对每个物品i计算兴趣分数 s_i和多样性分数 d_i

（4）根据 $s_i + d_i$对剩余4800个物品排序，分数最高的300个物品进入精排

2.3.3 召回多样性

双塔模型

（1）添加噪声

用户塔将用户特征作为输入，输出用户的向量表征；然后在向量数据库中做ANN检索，召回向量相似度高的物品

线上做召回时（计算出用户向量之后，做ANN检索之前），往用户向量中添加随机噪声

用户的兴趣越窄（比如用户最近交互的n个物品只覆盖少数几个类目），则添加的噪声越强

添加噪声使召回物品更多样，可以提升推荐系统核心指标

（2）抽样用户行为序列

步骤：

a. 将用户最近交互的n个物品（用户行为序列）作为用户塔的输入

b. 保留最近的r个物品（r远小于n）

c. 从剩余的n-r个物品中随机抽样t个样品（t远小于n），可以是均匀抽样，也可以是非均匀抽样让类目平衡

d. 将得到的r+t个物品作为用户行为序列，而不是用全部n个物品

抽样用户行为序列为什么能涨指标?

一方面，注入随机性，使召回结果更多样化

另一方面，n可以非常大，可以使召回结果覆盖到用户很久以前的兴趣

U2I2I：抽样用户行为序列

U2I2I（user→item→item）中的第一个item是指用户**最近交互**的n个物品之一，在U2I2I中叫作种子物品

问题：这n个物品覆盖的类目数可能较少，且类目不平衡

系统共有200个类目，某用户的n个物品只覆盖了15个

例如：足球类目的物品有0.4n个，电视剧类目的物品有0.2n个，其余类目的物品数均少于0.05n个

做法：非均匀**随机抽样**，从n个物品中选出t个，让类目平衡（和双塔模型中该操作类似），将这t个物品作为U2I2I的种子物品

一方面，类目更平衡，多样性更好；另一方面，n可以更大，覆盖的类目更多

2.3.4 探索流量

每个用户曝光的物品中有2%是非个性化的，做兴趣探索

具体做法：

（1）维护一个精选内容池，其中物品均为交互率指标高的优质物品（内容池可以分人群，比如30-40岁男性内容池）

（2）从精选内容池中随机抽样几个物品，跳过排序，直接插入（否则容易因不符合兴趣点被淘汰）最终排序结果

做法依据：缺少了用户的个性化，就要通过提升物品质量来吸引用户，用高质量弥补缺少个性化带来的损失

兴趣探索在短期内负向影响核心指标，但长期会产生正向影响

2.4 特殊对待特殊人群

原因：

（1）新用户、低活用户的行为很少，个性化推荐不准确

（2）新用户、低活用户容易流失，要想办法促使他们留存

（3）特殊用户的行为（如点击率、交互率）不同于主流用户，基于全体用户行为训练出的模型在特殊用户人群上有偏差

方式（以下方式仅针对特殊人群）：

2.4.1 构造特殊内容池

特殊内容池的构建：用于特殊用户人群的召回

在个性化召回不准确的情况下，保证内容质量好是关键

针对特定人群的特点构造特殊内容池，能提升用户满意度（如：对喜欢留评论的中年女性，构造促评论内容池，满足这些用户的互动需求）

方法1：根据物品获得的交互次数、交互率选择优质物品

圈定人群：只考虑特定人群，例如18-25岁一二线城市男性

构造内容池：用该人群对物品的交互次数、交互率给物品打分，选出分数最高的物品进入内容池

*由于范围是指定类别的人群，内容池有弱个性化的效果

*内容池需要定期更新加入新物品，排除交互率低和失去时效性的老物品

方法2：做因果推断，判断物品对人群留存率的贡献，根据贡献值选物品

特殊内容池的召回：通常使用双塔模型从特殊内容池中召回

双塔模型是个性化的，但对新用户而言双塔模型的个性化做不准，于是需要靠高质量内容、弱个性化做弥补

额外的训练代价：

对于正常用户，不论有多少内容池，只训练一个双塔模型

对于新用户，由于历史交互记录很少，需要单独训练模型

额外的推理代价：

内容池定期更新，然后要更新ANN索引

线上做召回时，需要做ANN检索

特殊内容池都很小，所以需要的额外算力不大

2.4.2 使用特殊排序策略

排除低质量物品

对于新用户、低活用户这类特殊人群，业务上只关注留存，不在乎消费（总曝光量、广告收入、电商收入）

对于新用户、低活用户，要少出广告甚至不出广告

新发布的物品不在新用户、低活用户上做探索；只在活跃的老用户上探索，对新物品提权（boost）

原因：物品新发布时，推荐做的不准，会损害新用户、低活用户的体验

差异化融分公式

新用户、低活用户的点击交互行为不同于正常用户——

低活用户的人均点击量很小，没有点击就不会有进一步的交互

因此，在低活用户的融分公式中，相较于普通用户，需要提高预估点击率的权重，或保留几个曝光坑位给预估点击率最高的几个物品

例如：精排从500个物品中选50个作为推荐结果，其中3个坑位给点击率最高的物品，剩余47个坑位由融分公式决定

（甚至可以把点击率最高的物品直接排在第一，确保用户一定能看到）

2.4.3 使用特殊排序模型

排序模型是拿全体用户训练的，给特殊用户的预估不准

例如：如果一个APP的用户90%都是女性，用全体用户数据训练出的模型，对男性用户做的预估就有偏差

问题：对于特殊用户，如何让排序模型预估更准？

方法1：大模型+小模型

a. 用全体用户行为训练大模型，大模型的预估p拟合用户行为y
b. 用特殊用户的行为训练小模型，小模型的预估q拟合大模型的残差y-p（对大模型起到纠正作用）
c. 对主流用户只用大模型做预估p
d. 对特殊用户，结合大模型和小模型的预估p+q

方法2：融合多个experts，类似MMoE

只用一个模型，模型有多个experts，各输出一个向量

对experts的输出做加权平均，权重仅根据用户特征计算得到（和第三章MMoE不同）

以新用户为例，模型将用户的新老、活跃度等特征作为输入，输出权重，用于对experts做加权平均

方法3：大模型预估之后，对小模型做校准

a. 用大模型预估点击率、交互率

b. 将用户特征、大模型预估点击率和交互率作为小模型（例如GBDT）的输入

c. 在特殊用户人群的数据上训练小模型，小模型的输出拟合用户真实行为（纠正大模型在特殊用户上的偏差）

（不建议每个用户人群使用一个排序模型，推荐系统同时维护多个大模型，这样维护代价太大）

2.5 利用交互行为

用户的交互行为：点赞、收藏、转发、关注、评论 .....

利用交互行为的方法：将模型预估的交互率用于排序

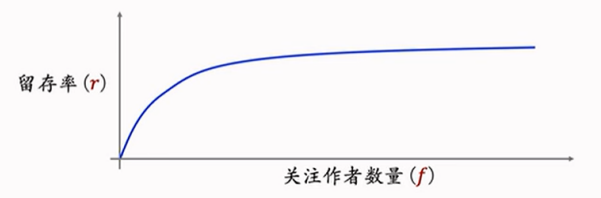
模型将交互行为当做预估的目标，将预估的点击率、交互率做融合，作为排序的依据

2.5.1 关注

关注量对留存的价值

对于一位用户，他关注的作者越多，则平台对它的吸引力越强

用户留存率（r）与他关注的作者数量（f）正相关



如果f较小，推荐系统需要促使用户关注更多作者

如何利用关注关系提升用户留存?

方法1：用排序策略提升关注量

a. 对于用户u，模型预估候选物品i的关注率为p_i

b. 设用户u已经关注了f个作者

c. 定义单调递减函数w(f)，用户已经关注的作者越多，则w(f)越小

d. 在排序融分公式中添加 $w(f)\cdot p_i$，用于促关注，从而给物品i带来很大加分，在用户关注作者数量很少时，更大可能关注物品i的作者

方法2：构造促关注内容池和召回通道

这个内容池中物品的关注率高，可以促关注

如果用户关注的作者数f较小，则对该用户使用该内容池

召回配额可以固定，也可以与f负相关

粉丝数对促发布的价值

UGC平台将作者发布量、发布率作为核心指标，希望作者多发布

作者发布的物品被平台推送给用户，会产生点赞、评论、关注等交互，其中关注、评论的交互可以提升作者发布积极性

作者粉丝数越少，则每增加一个粉丝对发布积极性的提升越大

具体方式：用排序策略帮助低分新作者涨粉

a. 某作者a的粉丝数（被关注数）为 f_a

b. 作者a发布的物品i可能被推荐给用户u，模型预估关注率为 p_{ui}

c. 定义单调递减函数 $w(f_a)$作为权重；作者a的粉丝越多，则 $w(f_a)$越小，给作者不会带来太多激励

d. 在排序融分公式中添加 $w(f_a)\cdot p_{ui}$，帮助低粉作者涨粉

隐式关注关系

召回通道U2A2I：user→author→item

显式关注关系：用户u关注了作者a，将a发布的物品推荐给u（点击率、交互率通常高于其它召回通道）

隐式关注关系：用户u喜欢看作者a发布的物品，但是u没有关注a

隐式关注的作者数量远大于显式关注，挖掘隐式关注关系，构造U2A2I召回通道，可以提升推荐系统核心指标

2.5.2 转发（分享）

A平台用户将物品转发到B平台，可以为A吸引站外流量

推荐系统做促转发（也叫分享回流）可以提升DAU和消费指标

简单提升转发次数是否有效?

模型预估转发率为p，融分公式中有一项w·p，让转发率大的物品更容易获得曝光机会

增大权重w可以促转发，吸引站外流量，但是会负面影响点击率和其他交互率（并不是转发到的地方的用户都对此满意）

KOL建模

目标：在不损害点击和其他交互的前提下，尽量多吸引站外流量

其他平台的Key Opinion Leader（KOL，即大V）的转发，可以吸引大量站外流量

注意是"其他平台"，这样即使他在站内没有粉丝，转发价值依然很大，因为他在其他平台有很强的吸引力和号召力

举例：我在抖音上有2w粉丝，微博10个粉丝，我把微博的内容转发到抖音，会吸引很多流量，但反过来就不行

如何判断本平台的用户是不是其他平台的KOL?

考察该用户历史上的转发能带来多少站外流量（例如转发到抖音的流量高，则有理由判断他可能是抖音的KOL）

识别出的站外KOL之后，该如何用于排序和召回?

方法1：排序融分公式中添加额外的一项 $k_u \cdot p_{ui}$

k_u：如果用户u是站外KOL，则它的值大

p_{ui}：用户推荐物品i模型预估的转发率

如果u是站外KOL，则多给他曝光他可能转发的物品，让本平台的信息流量更多，提升大盘指标

方法2：构造促转发内容池和召回通道，对站外KOL生效

2.5.3 评论

评论的发布价值:

促使新物品获得评论，提升作者发布积极性

如果新发布物品尚未获得很多评论，则给预估评论率提权，让物品尽快获得评论

排序融分公式中添加额外一项 $w_i \cdot p_i$

w_i：权重，与物品i已有的评论数量负相关

p_i：用户推荐物品i模型预估的评论率

评论的留存价值：

有的用户喜欢留评论，喜欢跟作者、评论区用户互动

给这样的用户添加促评论的内容池，让他们更多机会参与讨论，这样有利于提升这些用户的留存

鼓励高质量评论用户多留评论：

有的用户常留高质量评论（评论点赞量高）

高质量评论对作者、其他用户的留存有贡献（作者、其他用户觉得这样的评论有趣或有帮助）

推荐系统用排序和召回策略鼓励高质量评论用户多留评论