

双塔模型： 正负样本

王树森

ShusenWang@xiaohongshu.com

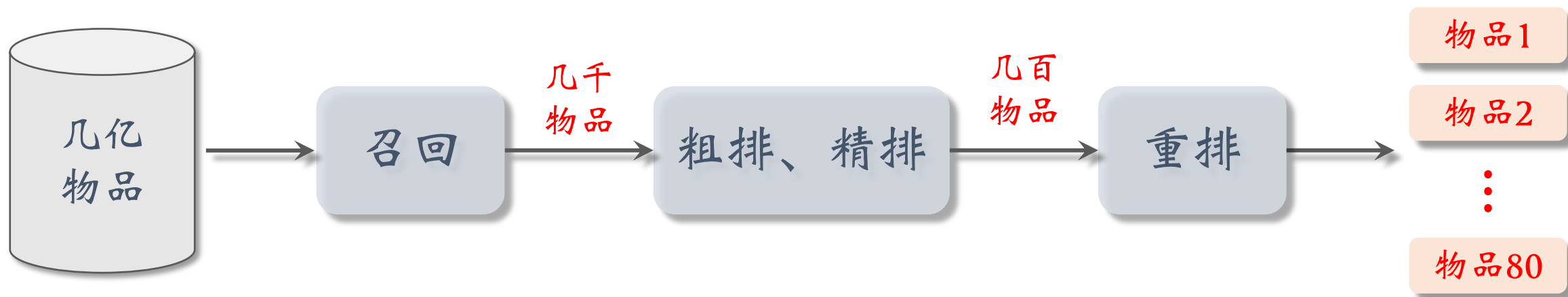
<http://wangshusen.github.io/>



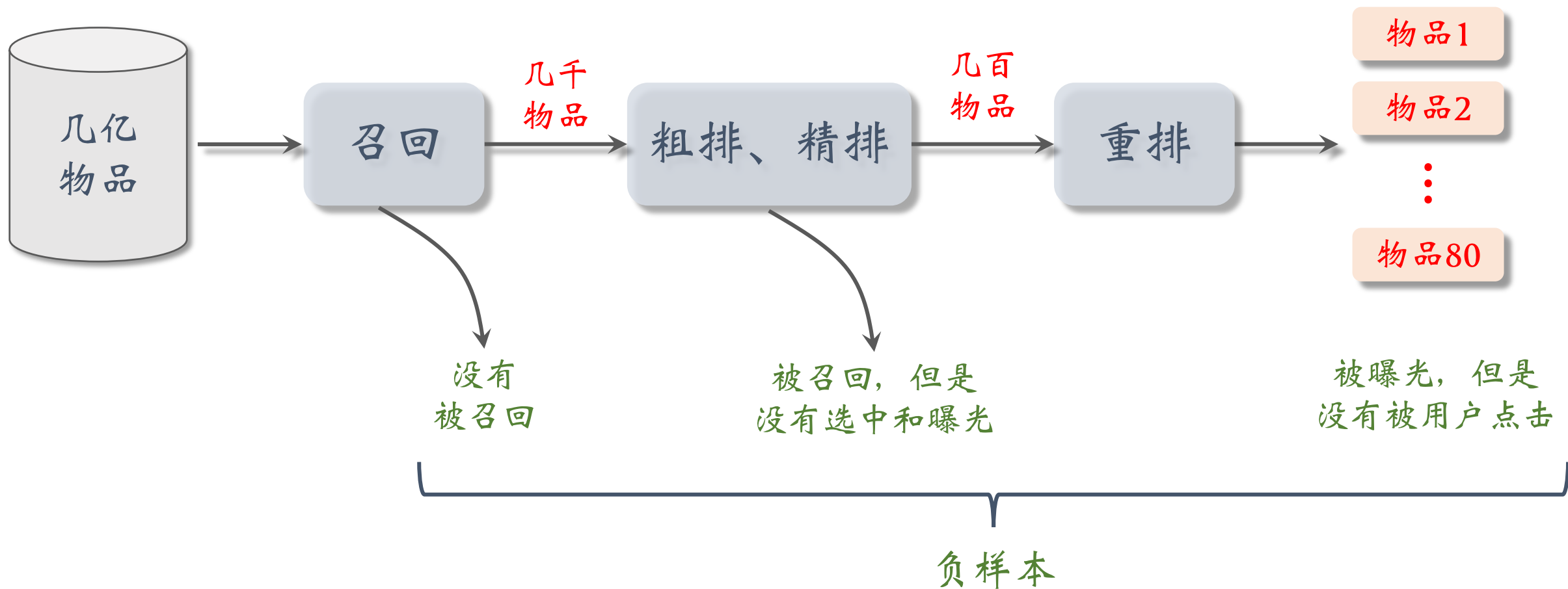
正样本

- 正样本：曝光而且有点击的 **用户—物品** 二元组。
(用户对物品感兴趣)。
- 问题：少部分物品占据大部分点击，导致正样本大多是热门物品。
- 解决方案：过采样冷门物品，或降采样热门物品。
 - 过采样 (up-sampling)：一个样本出现多次。
 - 降采样 (down-sampling)：一些样本被抛弃。

推荐系统的链路



如何选择负样本？



简单负样本

简单负样本：全体物品

- 未被召回的物品，大概率是用户不感兴趣的。
- 未被召回的物品 \approx 全体物品
- 从全体物品中做抽样，作为负样本。
- 均匀抽样 or 非均匀抽样？

简单负样本：全体物品

均匀抽样：对冷门物品不公平

- 正样本大多是热门物品。
- 如果均匀抽样产生负样本，负样本大多是冷门物品。

非均抽采样：目的是打压热门物品

- 负样本抽样概率与热门程度（点击次数）正相关。
- 抽样概率 $\propto (\text{点击次数})^{0.75}$ 。

简单负样本：Batch内负样本

用户：

物品：



点击



点击



⋮

⋮



点击



简单负样本：Batch内负样本

用户：

物品：



点击



点击



正样本

⋮

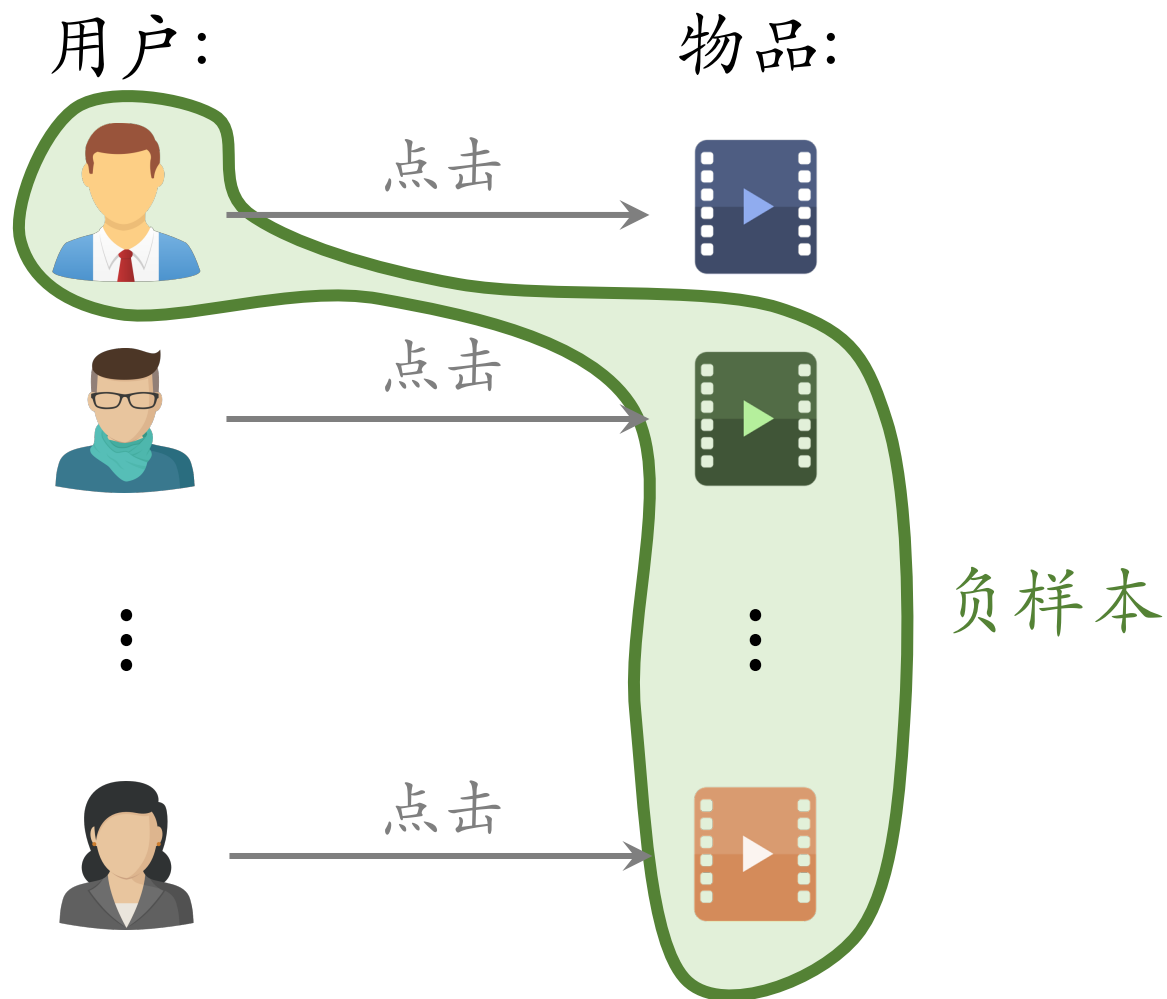
⋮



点击

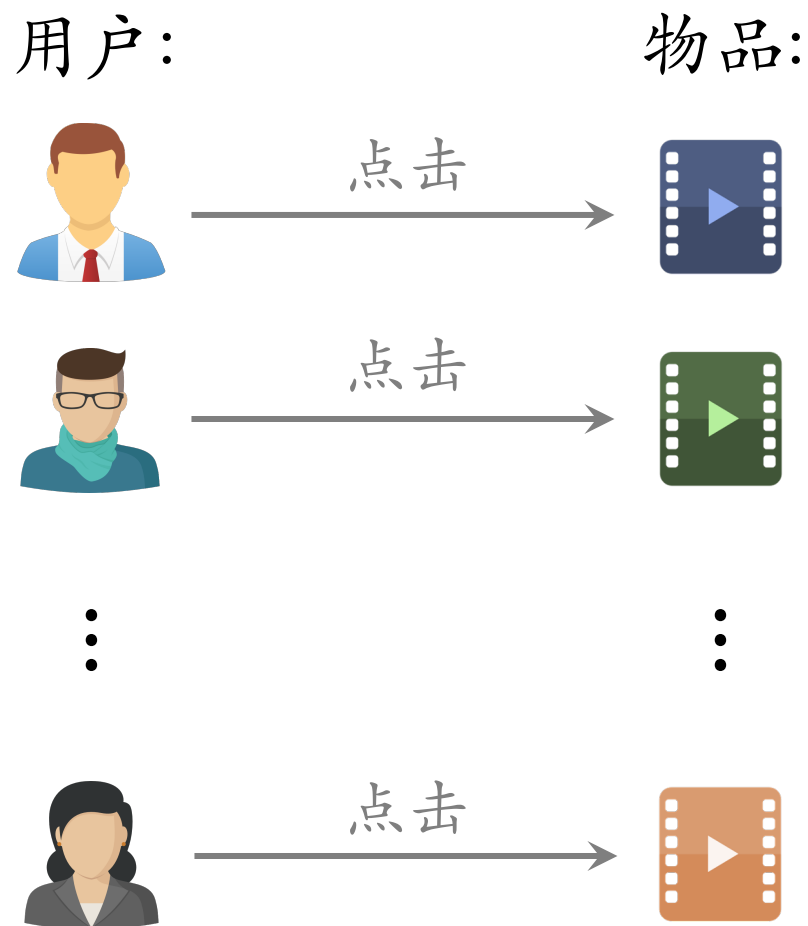


简单负样本: Batch内负样本



- 一个 batch 内有 n 个正样本。
- 一个用户和 $n - 1$ 个物品组成负样本。
- 这个 batch 内一共有 $n(n - 1)$ 个负样本。
- 都是简单负样本。（因为第一个用户不喜欢第二个物品。）

简单负样本：Batch内负样本



- 一个物品出现在 batch 内的概率 \propto 点击次数。
- 物品成为负样本的概率本该是 $\propto (\text{点击次数})^{0.75}$ ，但在这里是 \propto 点击次数。
- 热门物品成为负样本的概率过大。

参考文献：

- Xinyang Yi et al. [Sampling-Bias-Corrected Neural Modeling for Large Corpus Item Recommendations](#). In *RecSys*, 2019.

简单负样本：Batch内负样本

用户：



点击



物品：



点击



⋮

⋮



点击



- 物品 i 被抽样到的概率：

$$p_i \propto \text{点击次数}$$

- 预估用户对物品 i 的兴趣：

$$\cos(\mathbf{a}, \mathbf{b}_i)$$

- 做训练的时候，调整为：

$$\cos(\mathbf{a}, \mathbf{b}_i) - \log p_i$$

参考文献：

- Xinyang Yi et al. [Sampling-Bias-Corrected Neural Modeling for Large Corpus Item Recommendations](#). In *RecSys*, 2019.

困难负样本

困难负样本

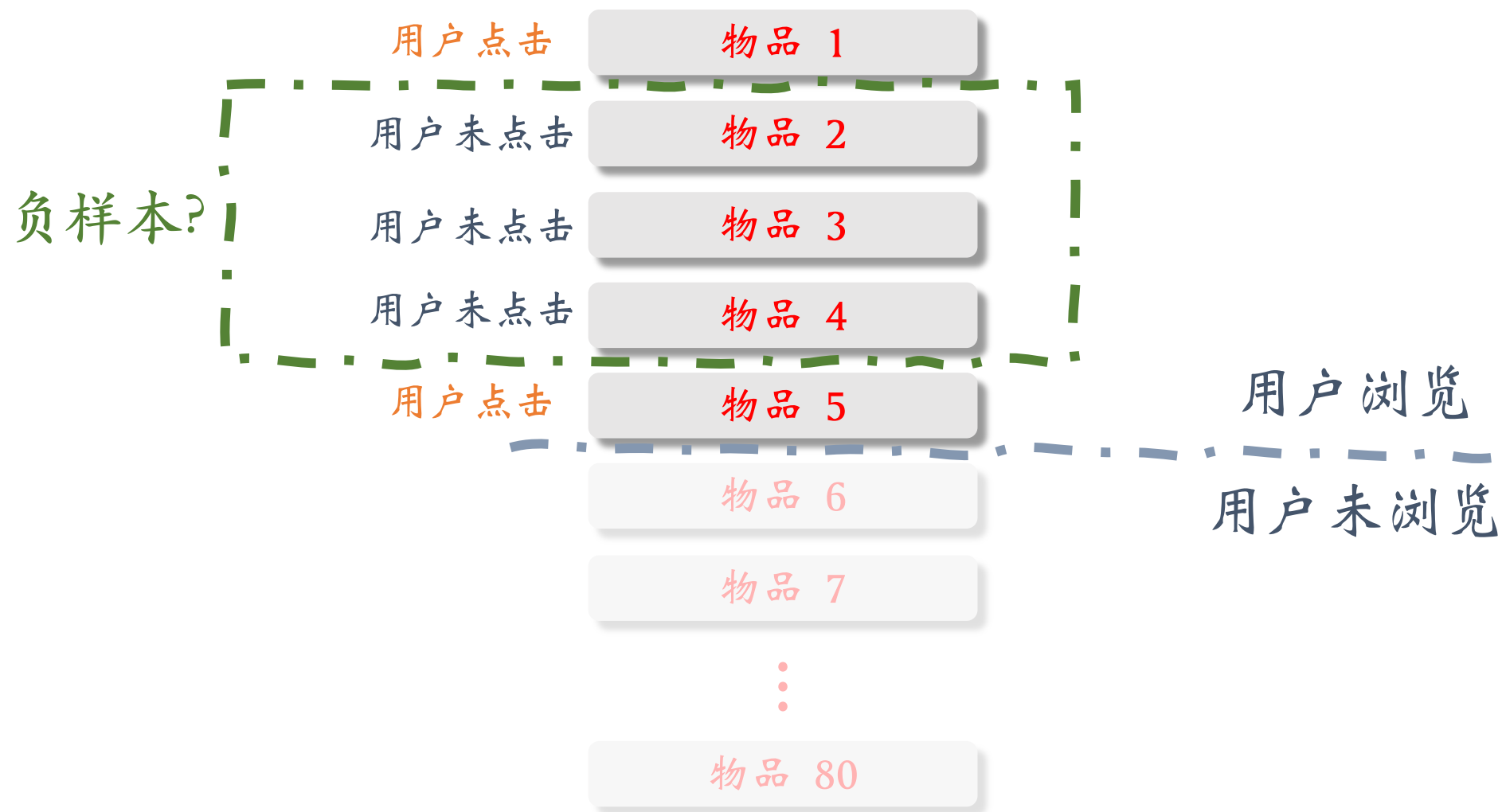
- 困难负样本：
 - 被粗排淘汰的物品（比较困难）。
 - 精排分数靠后的物品（非常困难）。
- 对正负样本做二元分类：
 - 全体物品（简单）分类准确率高。
 - 被粗排淘汰的物品（比较困难）容易分错。
 - 精排分数靠后的物品（非常困难）更容易分错。

训练数据

- 混合几种负样本。
- 50%的负样本是全体物品（简单负样本）。
- 50%的负样本是没通过排序的物品（困难负样本）。

常见的错误

曝光但是没有点击



曝光但是没有点击



训练召回模型不能用这类负样本

训练排序模型会用这类负样本

选择负样本的原理

召回的目标：快速找到用户可能感兴趣的物品。

- 全体物品（easy）：绝大多数是用户根本不感兴趣的。
- 被排序淘汰（hard）：用户可能感兴趣，但是不够感兴趣。
- 有曝光没点击（没用）：用户感兴趣，可能碰巧没有点击。

可以作为排序的负样本，
不能作为召回的负样本

总结

- ➡ • 正样本：曝光而且有点击。
- ➡ • 简单负样本：
 - 全体物品。
 - batch内负样本。
- ➡ • 困难负样本：被召回，但是被排序淘汰。
- ➡ • 错误：曝光、但是未点击的物品做召回的负样本。

Thank You!

<http://wangshusen.github.io/>

长期招聘优秀的算法工程师

- 部门：小红书社区技术部。
- 方向：搜索、推荐。
- 职位：校招、社招、实习。
- 地点：上海、北京。
- 联系方式：`ShusenWang@xiaohongshu.com`