物品冷启动：简单的召回通道

王树森

ShusenWang@xiaohongshu.com

http://wangshusen.github.io/

小红书

召回的难点

召回的依据

✓ 自带图片、文字、地点。

✓ 算法或人工标注的标签。

✗ 没有用户点击、点赞等信息。

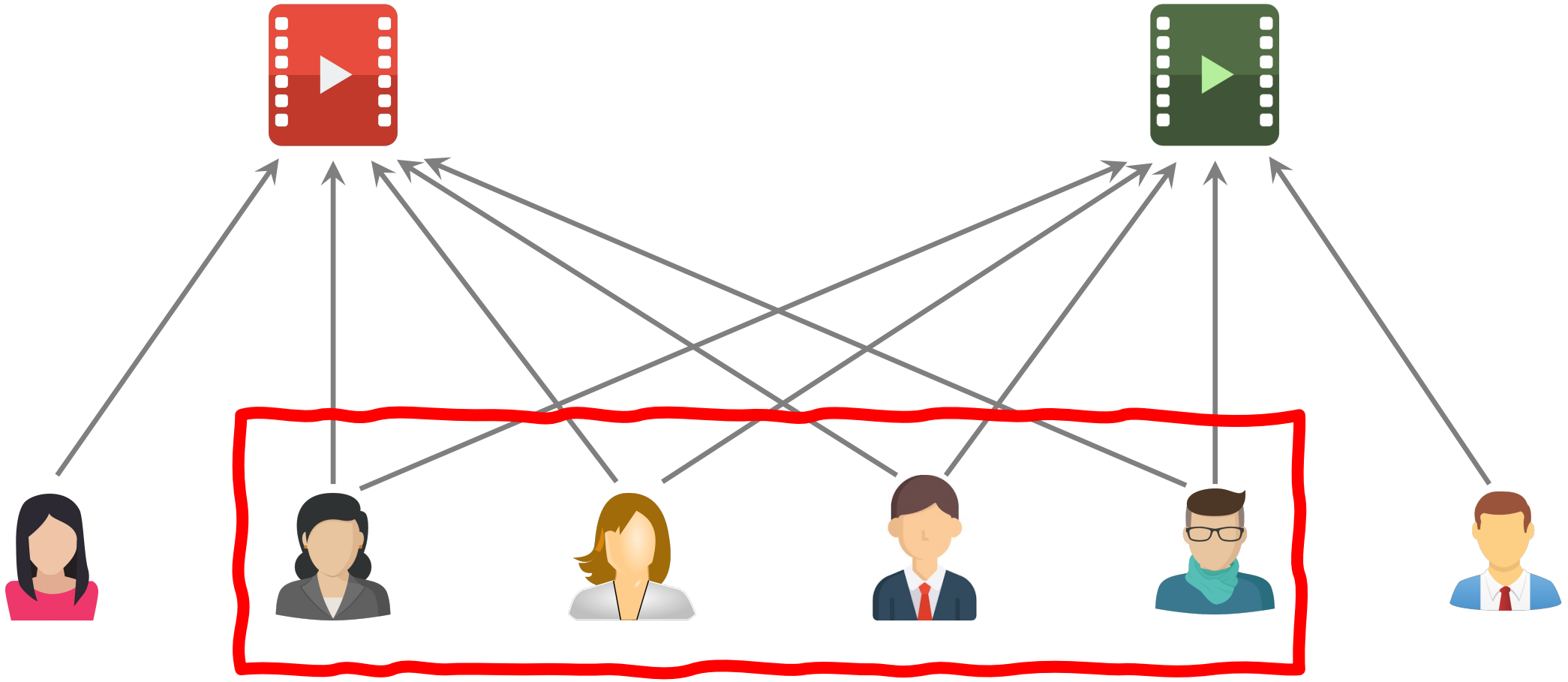
✗ 没有笔记 ID embedding。

冷启召回的困难

- 缺少用户交互，还没学好笔记 ID embedding，导致双塔模型效果不好。
- 缺少用户交互，导致 ItemCF 不适用。

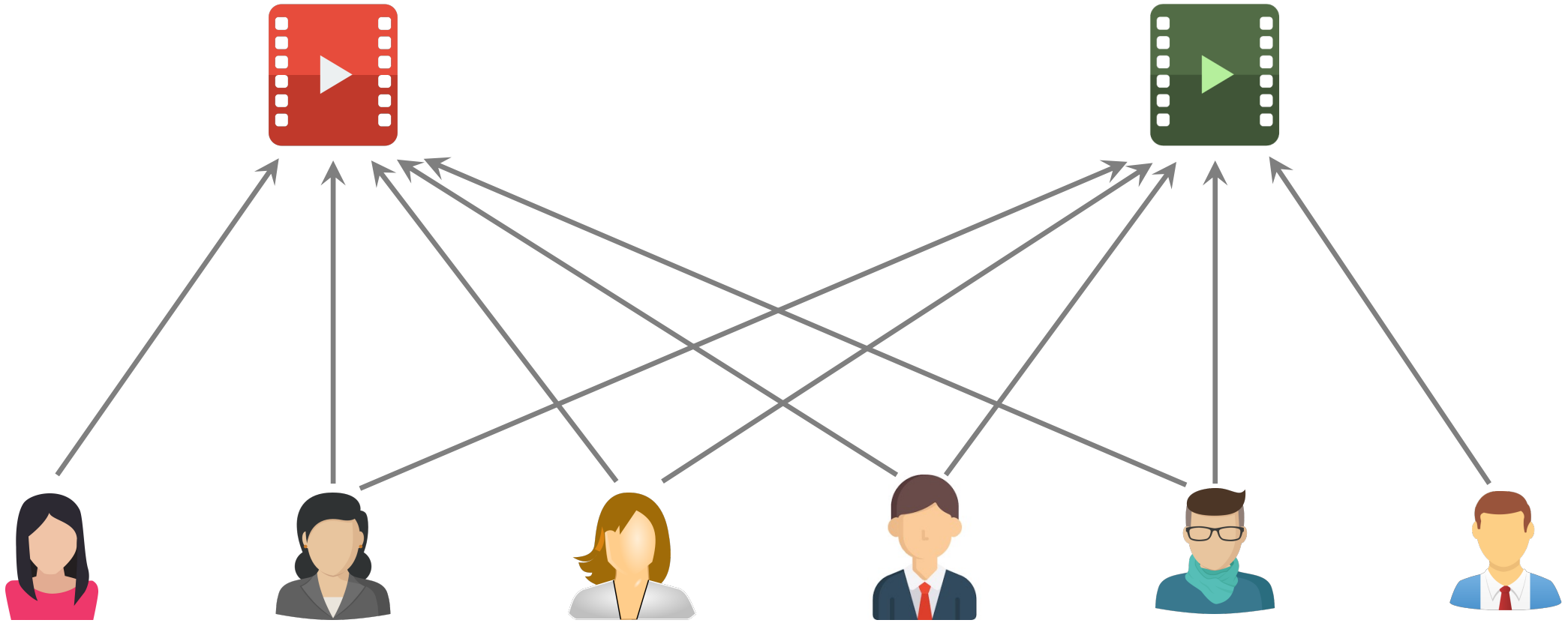
冷启召回的困难

ItemCF 不适用于物品冷启动



冷启召回的困难

ItemCF 不适用于物品冷启动

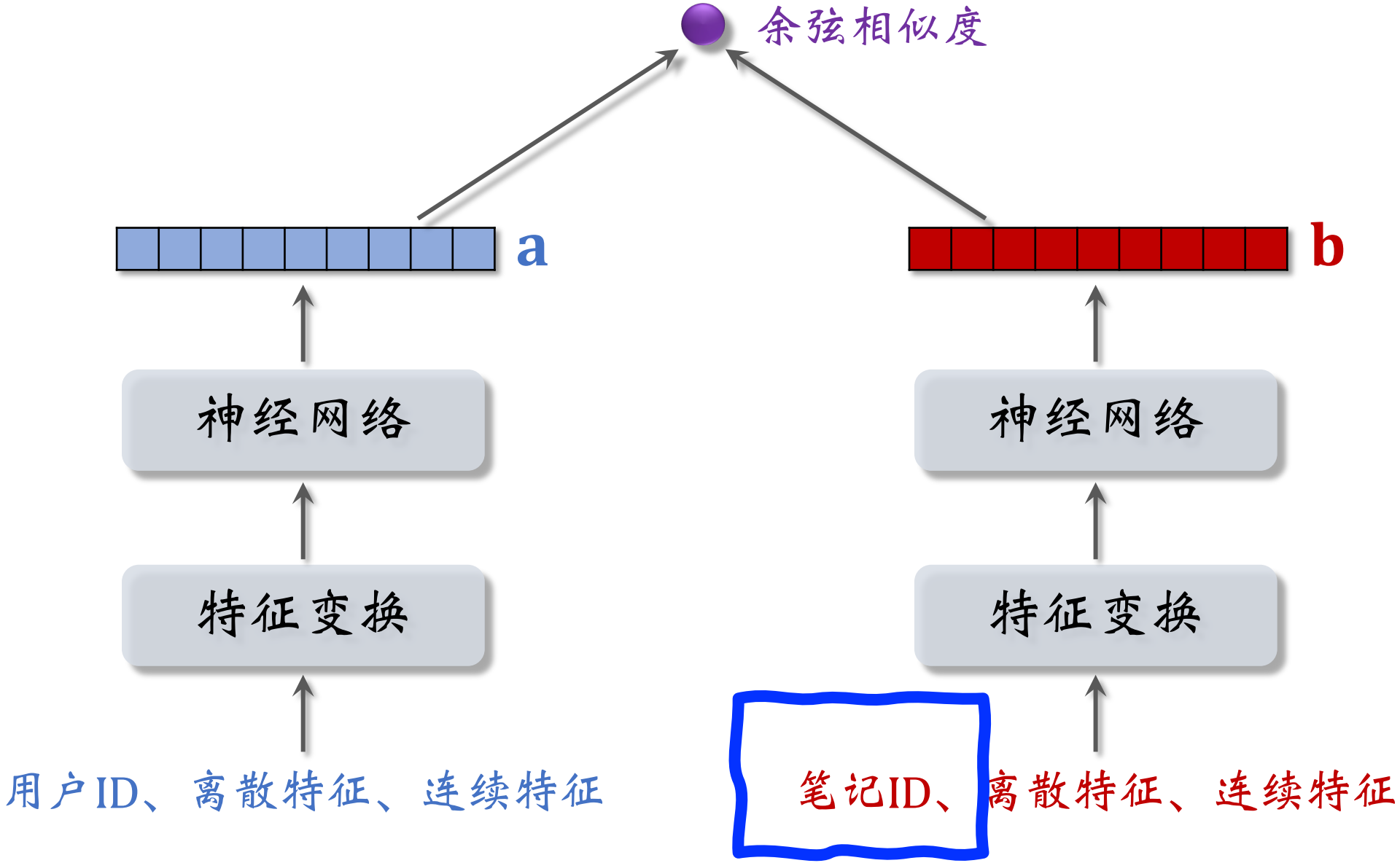


召回通道

- ✗ ItemCF召回（不适用）
- ? 双塔模型（改造后适用）
- ✓ 类目、关键词召回（适用）
- ✓ 聚类召回（适用）
- ✓ Look-Alike召回（适用）

双塔模型

双塔模型



ID Embedding

改进方案 1：新笔记使用 default embedding。

- 物品塔做 ID embedding 时，让所有新笔记共享一个 ID，而不是用自己真正的 ID。
- Default embedding：共享的 ID 对应的 embedding 向量。
- 到下次模型训练的时候，新笔记才有自己的 ID embedding 向量。

ID Embedding

改进方案 1：新笔记使用 default embedding。

改进方案 2：利用相似笔记 embedding 向量。

- 查找 top k 内容最相似的高曝笔记。
- 把 k 个高曝笔记的 embedding 向量取平均，作为新笔记的 embedding。

多个向量召回池

- 多个召回池，让新笔记有更多曝光机会。
  - 1 小时新笔记，
  - 6 小时新笔记，
  - 24 小时新笔记，
  - 30 天笔记。
- 共享同一个双塔模型，那么多个召回池不增加训练的代价。

类目召回

用户画像

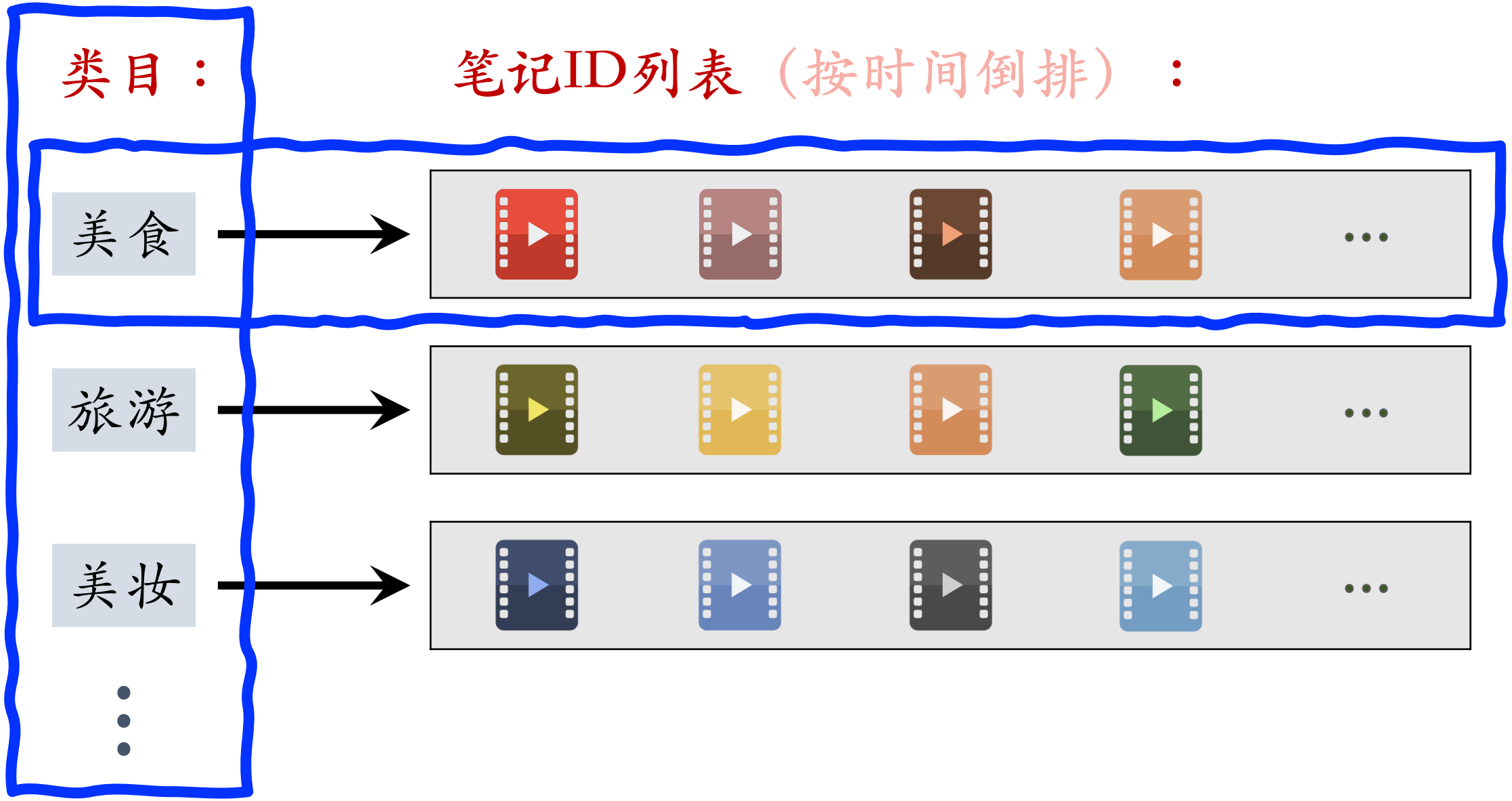
- 感兴趣的类目：

  美食、科技数码、电影……

- 感兴趣的关键词：

  纽约、职场、搞笑、程序员、大学……

基于类目的召回



基于类目的召回

- 系统维护类目索引：

  类目 → 笔记列表（按时间倒排）

- 用类目索引做召回：

  用户画像 → 类目 → 笔记列表

- 取回笔记列表上前 k 篇笔记（即最新的 k 篇）。

基于关键词的召回

- 系统维护关键词索引：

  关键词 → 笔记列表（按时间倒排）

- 根据用户画像上的关键词做召回。

缺点

- 缺点1：只对刚刚发布的新笔记有效。
  - 取回某类目/关键词下最新的 k 篇笔记。
  - 发布几小时之后，就再没有机会被召回。
- 缺点2：弱个性化，不够精准。

总结

✗ ItemCF召回（不适用）

? 双塔模型（改造后适用）

✓ 类目、关键词召回（适用）

✓ 聚类召回（适用）

✓ Look-Alike召回（适用）

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

小红书