

排序模型的特征

王树森

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



用户画像 (User Profile)

- 用户 ID (在召回、排序中做 embedding) 。
- 人口统计学属性：性别、年龄。
- 账号信息：新老、活跃度……
- 感兴趣的类目、关键词、品牌。

物品画像 (Item Profile)

- 物品 ID (在召回、排序中做 embedding) 。
- 发布时间 (或者年龄) 。
- GeoHash (经纬度编码) 、所在城市 。
- 标题、类目、关键词、品牌……
- 字数、图片数、视频清晰度、标签数……
- 内容信息量、图片美学……

用户统计特征

- 用户最近30天（7天、1天、1小时）的曝光数、点击数、点赞数、收藏数……
- 按照笔记**图文/视频**分桶。（比如最近7天，该用户对图文笔记的点击率、对视频笔记的点击率。）
- 按照笔记类目分桶。（比如最近30天，用户对美妆笔记的点击率、对美食笔记的点击率、对科技数码笔记的点击率。）

笔记统计特征

- 笔记最近30天（7天、1天、1小时）的曝光数、点击数、点赞数、收藏数……
- 按照用户性别分桶、按照用户年龄分桶……
- 作者特征：
 - 发布笔记数
 - 粉丝数
 - 消费指标（曝光数、点击数、点赞数、收藏数）

场景特征 (Context)

- 用户定位 GeoHash (经纬度编码)、城市。
- 当前时刻 (分段, 做 embedding)。
- 是否是周末、是否是节假日。
- 手机品牌、手机型号、操作系统。

特征处理

- 离散特征：做 embedding。
 - 用户ID、笔记ID、作者ID。
 - 类目、关键词、城市、手机品牌。
- 连续特征：做分桶，变成离散特征。
 - 年龄、笔记字数、视频长度。
- 连续特征：其他变换。
 - 曝光数、点击数、点赞数等数值做 $\log(1 + x)$ 。
 - 转化为点击率、点赞率等值，并做平滑。

小结

1. 用户画像特征。
2. 笔记画像特征。
3. 用户统计特征。
4. 笔记统计特征。
5. 场景特征。

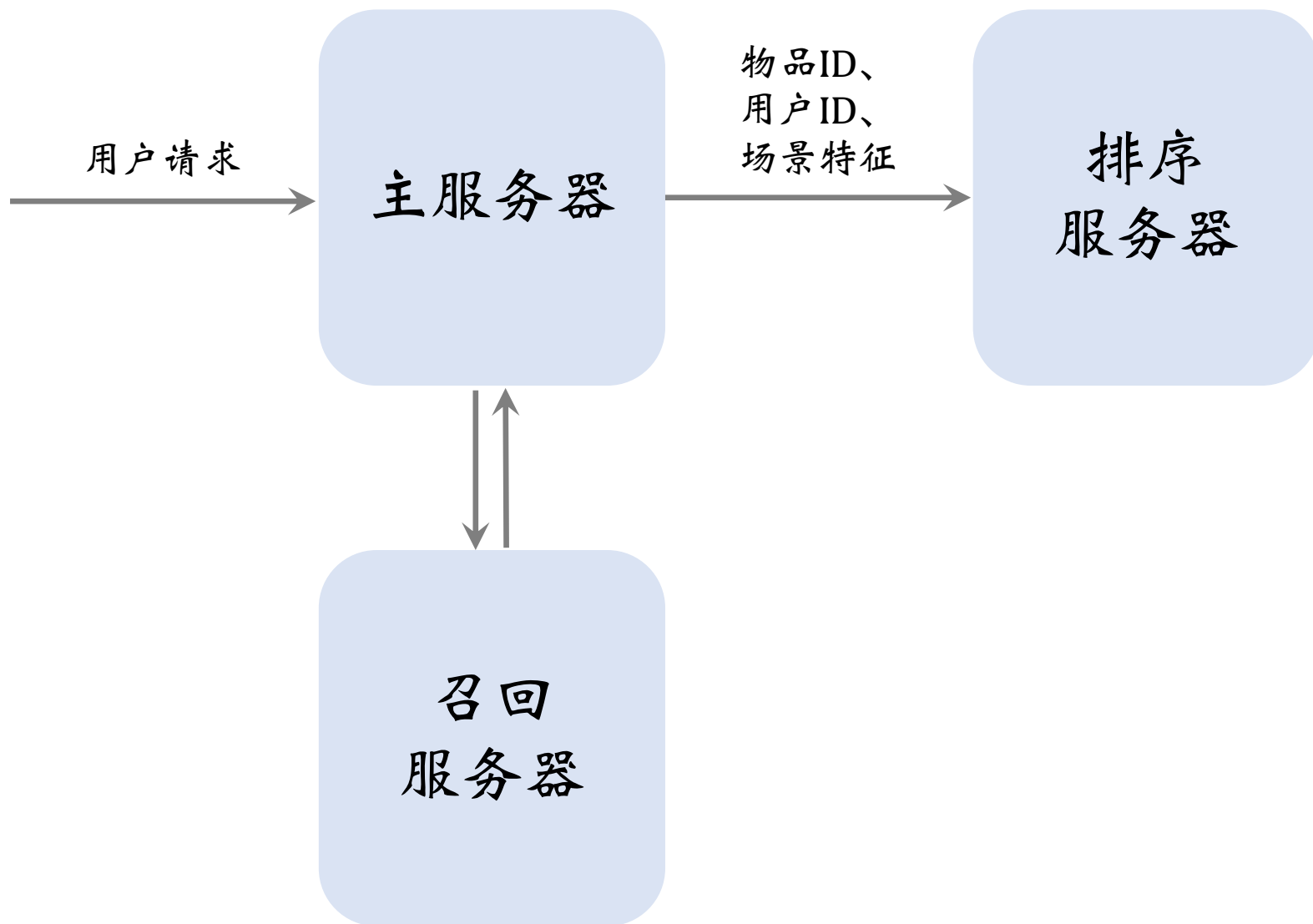
特征覆盖率

- 很多特征无法覆盖 100% 样本。
- 例：很多用户不填年龄，因此用户年龄特征的覆盖率远小于 100%。
- 例：很多用户设置隐私权限，APP 不能获得用户地理定位，因此场景特征有缺失。
- 提高特征覆盖率，可以让精排模型更准。

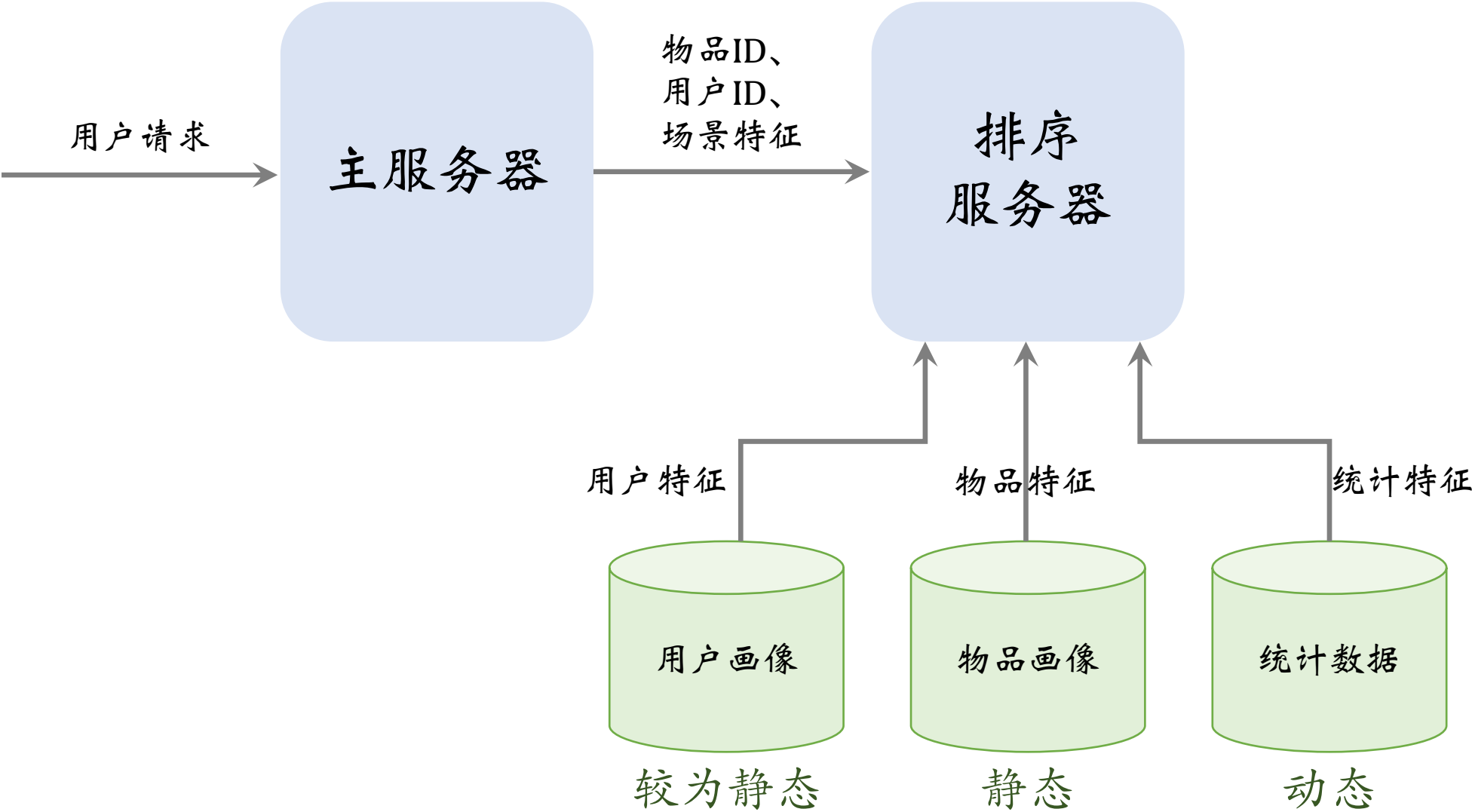
数据服务

1. 用户画像 (User Profile) ◦
2. 物品画像 (Item Profile) ◦
3. 统计数据 ◦

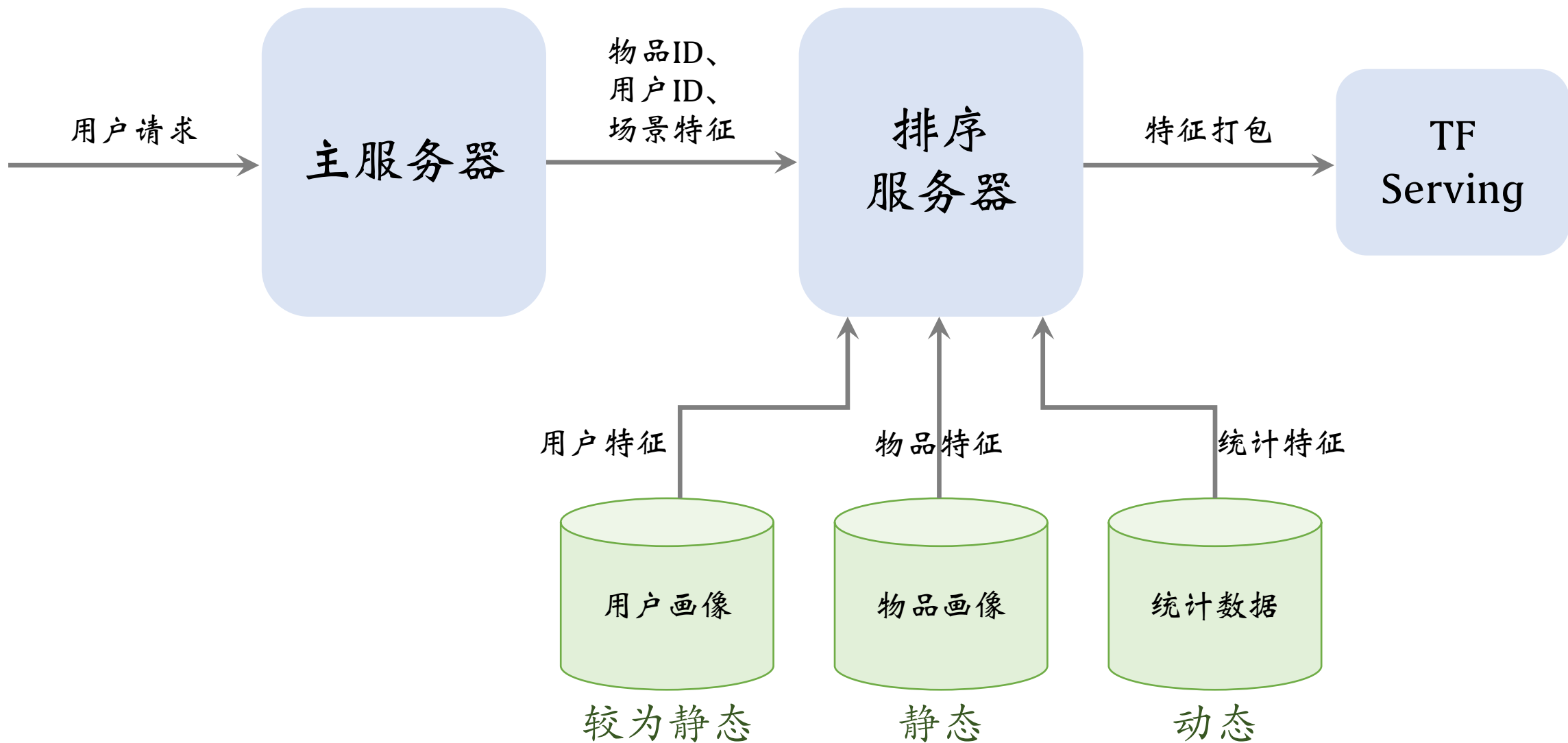
数据服务



数据服务



数据服务



Thank You!

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