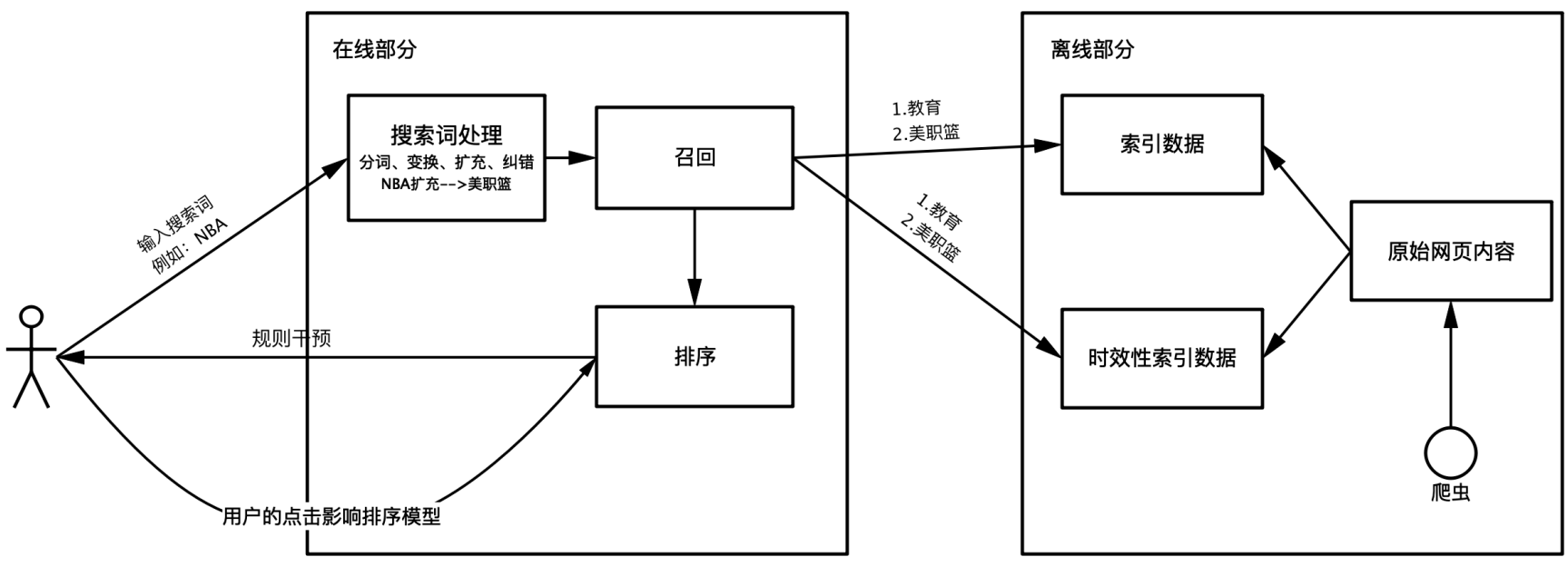
《内容算法》笔记--part1

2019年4月7日 星期日　　下午2:58

→ 架构

▶ 搜索系统架构



▶ 推荐系统架构

- 相同点：

  信息与用户意图之间的匹配

- 不同点：

- 表意明确的查询词 vs 没有明确表达的偏好
- 用户行为-->影响内容价值评判 vs 用户行为-->影响内容价值评判+自身画像建立

→ 推荐的起点：断物识人

▶ 断物

标签		分类		聚类	
在不同的应用场景下，我们对标签全集进行有针对性的投射，用不同的标签以换取信息匹配效率的最大化		树状的，自上而下，每个节点都有严格的继承关系，兄弟节点具有可以被完全枚举的属性值		不下定义。基于某一维度的特征将相关物品组成一个集合，并告诉你这个新的物品同哪个集合相似	
权威性弱、 灵活性强、 完备性强		权威性强、 灵活性弱、 完备性弱			
PGC	UGC 需要经过清洗和归一处理				
	1.五星评价 2.标签输入 3.简短评论				
先基于产品场景快速覆盖主要标签 ——结合使用频次、专家建议——> 将部分入口收敛到树状的分类体系					

▶ 识人

- 通过标签来描述一个用户的特征集合
- 应用场景

- 精准广告营销
- 行业研究　eg.消费分析等
- 产品效率优化

- 数据来源

静态用户画像数据	动态用户画像数据	
独立于产品场景之外，有统计性意义	产品场景中，不同行为权重不同	
包含： 性别 学历 年龄 教育程度 婚育状况 常住位置（旅行者模块）	显式： 点赞 评论 （文本分析） 分享（以社会身份传递了立场态度，意义大） 关注 收藏 搜索 评分（根据历史平均分归一化）	隐式： 某页面的停留时长 用户的操作行为轨迹 播放比例/播放时长
	稀疏，权重更高	权重较低，补充验证
来源： 第三方联合登录 用户表单填写		

→ 推荐算法：物以类聚，人以群分

▶ 物以类聚：基于内容属性的相似性推荐

- 推荐与用户历史消费相似的新物品
- 相似度计算：

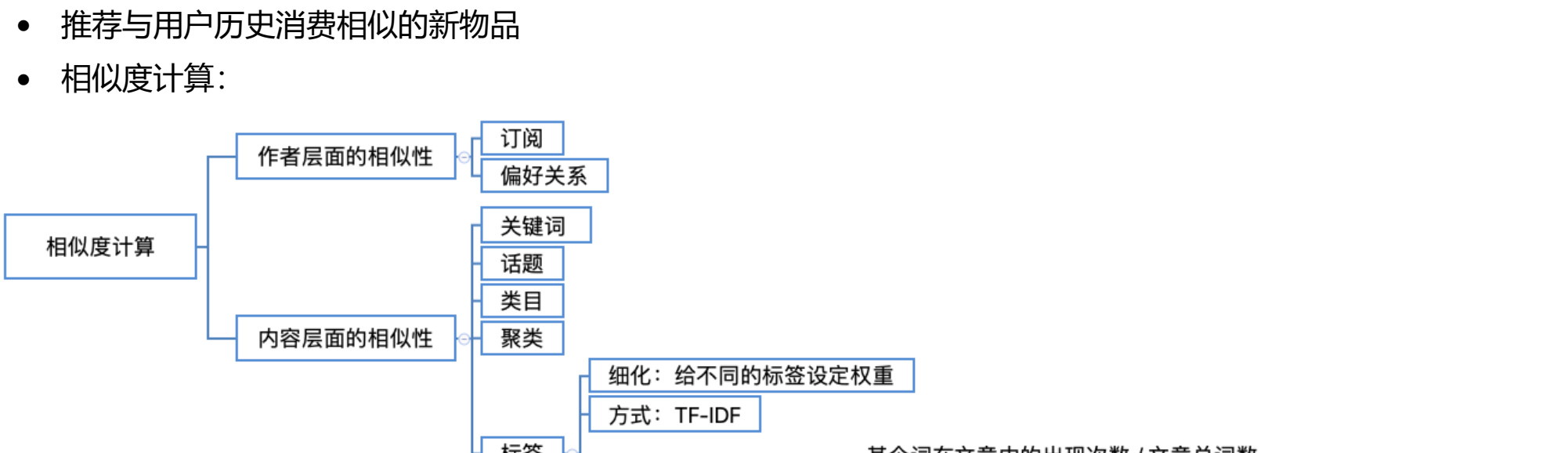
- 优点：

  只依赖于物品本身的特征，而不依赖于用户的行为，让新的物品、冷僻的物品都能得到展示的机会

- 缺点：

- 依赖于特征构建的完备性，存在一定成本
- 没有引入受众反馈因素

▶ 人以群分：基于用户行为的协同过滤



- 优点：

  不需要对物品或信息进行完整的标签化分析和建模

- 缺点：

- 领域无关，可以很好的发现用户的潜在兴趣爱好
- 依赖历史数据，新用户/新物品存在冷启动问题