智能推荐系统

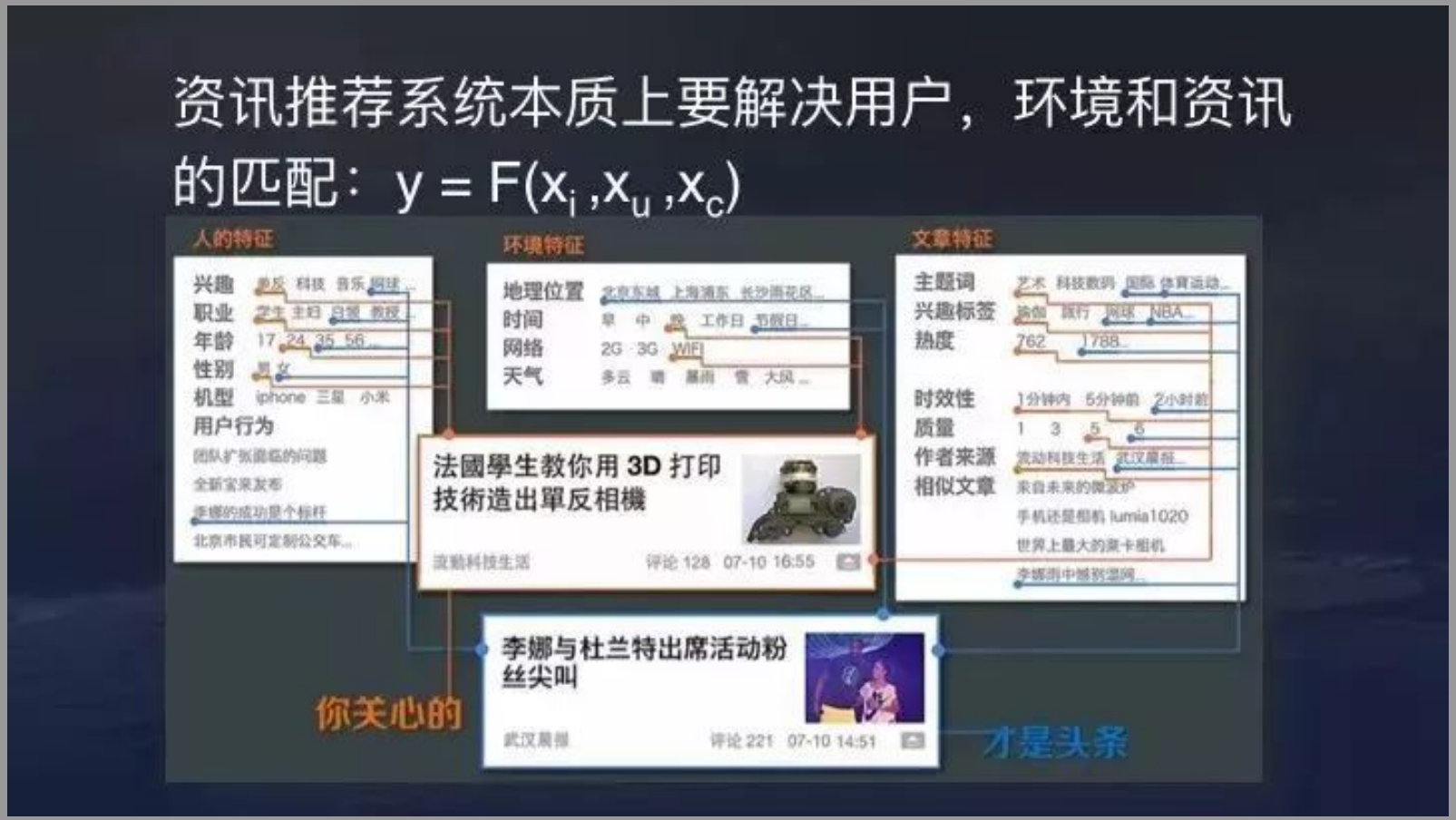
AII Media Literacy

闵勇

北京师范大学

April 18, 2022

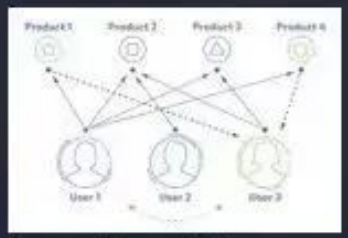
Outline


资讯推荐系统本质上要解决用户，环境和资讯的匹配：$y = F(x_i, x_u, x_c)$
人的特征
兴趣
职业
年龄
性别
机型
用户行为
环境特征
地理位置
时间
网络
天气
文章特征
主题词
兴趣标签
热度
时效性
质量
作者来源
相似文章
法國學生教你用 3D 打印技術造出單反相機
流動科技生活
评论 128 07-10 16:55
李娜与杜兰特出席活动粉丝尖叫
武汉晨报
评论 221 07-10 14:51
你关心的
才是头条

如何引入无法直接衡量的目标?

- 广告&特型内容频控
- 低俗内容打压&频控
- 标题党，低质，恶心内容打压
- 重要新闻置顶&强插&加权
- 低级别账号内容降权

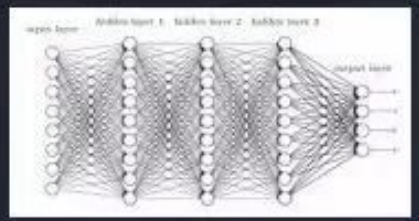
典型推荐算法



协同过滤

$$F(x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$$

Logistic Regression

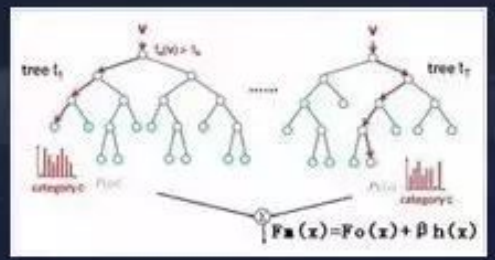


DNN

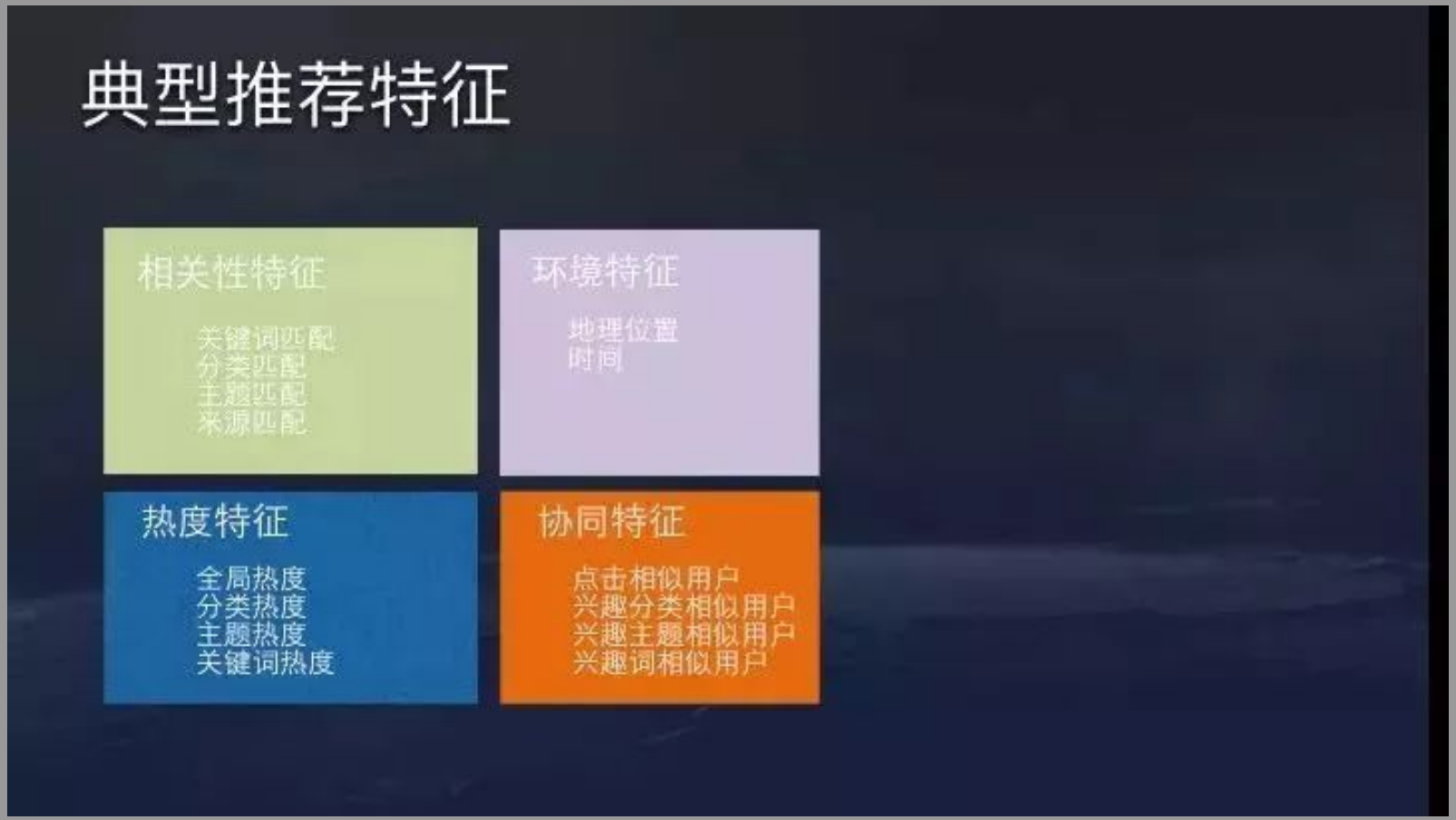
$$\hat{y}(x) = w_0 + \sum_{j=1}^{p} w_j x_j + \sum_{j=1}^{p} \sum_{j=j+1}^{p} x_j x_j \langle v_j v_j \rangle$$

$$\hat{y}(x) = w_0 + \sum_{j=1}^{p} w_j x_j + \sum_{j=1}^{p} \sum_{j=j+1}^{p} x_j x_j \sum_{f=1}^{k} v_{fj} v_{f,j}$$

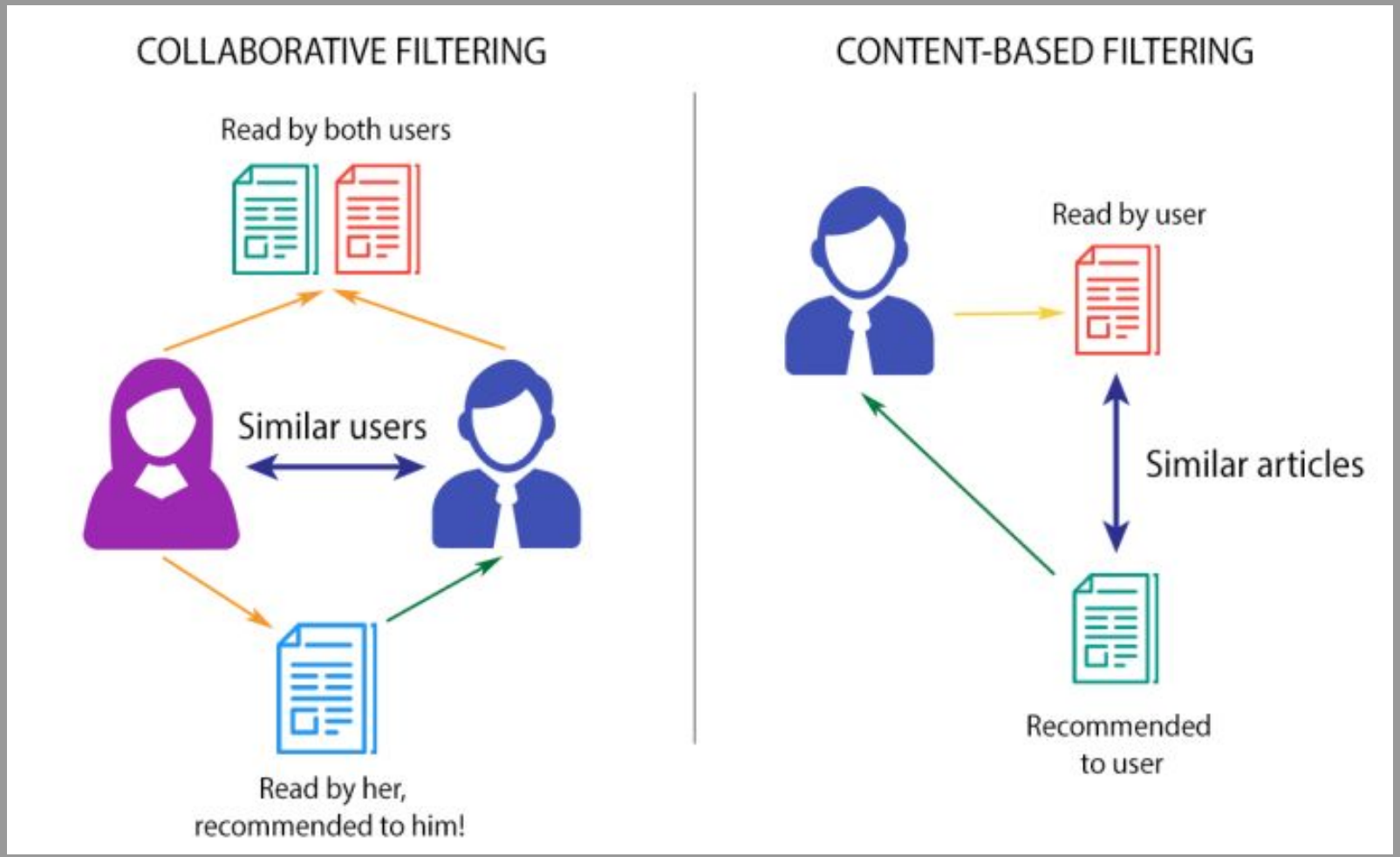
Factorization Machine



GBDT


典型推荐特征
相关性特征
关键词匹配
分类匹配
主题匹配
来源匹配
环境特征
地理位置
时间
热度特征
全局热度
分类热度
主题热度
关键词热度
协同特征
点击相似用户
兴趣分类相似用户
兴趣主题相似用户
兴趣词相似用户

协同过滤

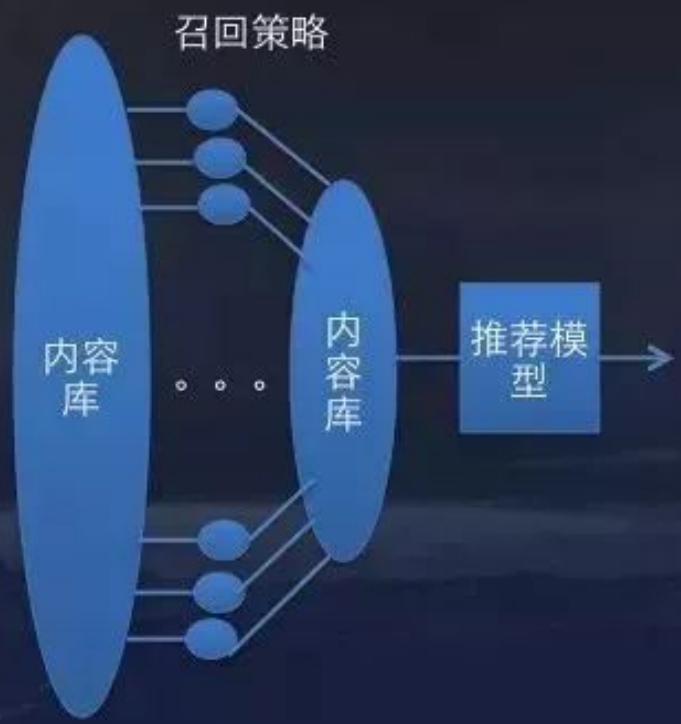

COLLABORATIVE FILTERING
Read by both users
Similar users
Read by her,
recommended to him!
CONTENT-BASED FILTERING
Read by user
Similar articles
Recommended
to user

大规模推荐模型的在线训练

- 用Storm集群实时处理样本数据（点击，展现，收藏，分享）
- 每收集一定量的用户数据就更新推荐模型
- 模型参数存储在高性能服务器集群，包含几百亿原始特征和数十亿向量特征

召回策略设计

- 推荐模型计算开销相对较大，完全依赖模型推荐成本过高
- 基于简化策略的召回模块可以有效平衡计算成本和效果




典型召回策略架构
用户兴趣标签
德甲 0.3
电商 0.2
O2O 0.2
搞笑 0.1
历史 0.1
军事 0.1
德甲
英超
互联网金融
搞笑
电商
离线更新tag倒排索引
根据兴趣标签拉取相应文章并rank top 结果
以兴趣分类召回为例，实际上这里的tag可以是各种显式兴趣标签和隐式兴趣特征

推荐系统的数据依赖

- 推荐模型的特征抽取需要用户侧和内容测的各种标签
- 召回策略需要获取用户侧和内容侧的各种标签
- 内容分析和用户标签挖掘是搭建推荐系统的基石

文本分析在推荐系统的应用

- 用户兴趣建模（user profile）：比如，给喜欢阅读【互联网】文章的用户打上【互联网】标签，给喜欢【小米】新闻的用户打上【小米】标签
- 帮助内容推荐：【魅族】的内容推荐给关心【魅族】的用户，【Dota】的内容推荐给关心【Dota】的用户
- 生成频道内容：【德甲】的内容进【德甲频道】，【瘦身】的内容进【瘦身频道】

文本特征case


查找文章: 4688699423 提交
4688699423 莎娃连续17次不敌小威 07-10 13:18 rate:18 展开>>
文章Profile
一级分类 展开>>
news_sports 2.5957
二级分类 展开>>
news_sports/tennis 0.7201
关键词2 展开>>
西班牙 0.9915 小威 0.9858 穆古拉扎 0.9845 女单决赛 0.9641
俄罗斯 0.9475 莎拉波娃 0.9282 莎娃 0.9208 小威康姆斯 0.9199
委内瑞拉 0.8738 锦标赛 0.7582 温网 0.6409 大满贯 0.5660
半决赛 0.4663
高亮关键词 展开>>
西班牙 0.9976 莎拉波娃 0.9886 俄罗斯 0.9856 小威康姆斯 0.9831
委内瑞拉 0.9823 小威 0.9498 穆古拉扎 0.9463 温网 0.9323
半决赛 0.7198 女单决赛 0.7114 大满贯 0.6948 波兰 0.6094

文本特征case

查找文章: 4688699423 提交

4688699423 莎娃连续17次不敌小威 07-10 13:18 rate:18 展开>>

文章Profile

2048Topic	展开>>
1233: 破发, 种子, 发球局, 发球, 影帅, 法网, 破发点, 首盘	0.7024
1464: 冠军, 夺冠, 决赛, 夺得, 奖杯, 问鼎, 赢得, 捧起	0.0755
887: 次数, 10次, 7次, 8次, 3次, 2次, 1次, 4次	0.0700
1485: 恐怖, 惊悚, 恐怖片, 吓人, 灵异, 诡异, 笔仙, 冷汗	0.0415
1822: 植物, 叶片, 产于, 果实, 栽培, 基部, 别名, 椭圆形	0.0353
1356: 时尚, 时装, 秀场, 设计师, 男装, 时装周, 时尚界, 模特	0.0305
1809: 拉美, 阿根廷, 委内瑞拉, 古巴, 南美, 墨西哥, 秘鲁, 智利	0.0207
229: 社会主义, 马克思主义, 革命, 资本主义, 马克思, 共产主义, 思想, 无产阶级	0.0186
1297: 法网, 纳达尔, 网球, 李娜, 费德勒, 大满贯, 温网, 红土	0.0055

新版实体词	展开>>
玛丽亚-莎拉波娃	0.9672
塞雷娜-威廉姆斯	0.9372
阿格涅什卡-拉德万斯卡	0.6391
温布尔登网球锦标赛	0.5021
法国网球公开赛	0.2950
委内瑞拉	0.2784
西班牙	0.1600
波兰	0.1485
俄罗斯	0.1237

文本特征对于推荐的独特价值

- 没有文本特征，推荐引擎无法工作
- 协同类特征无法解决文章冷启动问题
- 粒度越细的文本特征，冷启动能力越强 eg:
  【拜仁慕尼黑】 VS 【体育】

语义标签

特征	使用场景
分类	user profile；过滤频道内容；推荐召回；推荐特征
概念	过滤频道内容；标签搜索；推荐召回（Like）
实体	过滤频道内容；标签搜索；推荐召回（Like）

为什么分层?

- 每个层级粒度不一样，要求也有区别
- 分类体系要求覆盖全，希望任何一篇文章，总能找到合适的分类，精确性要求不高
- 实体体系不要求覆盖全，只要覆盖每个领域热门的人物，机构，作品，产品即可
- 概念体系负责表达比较精确，但是又属于抽象概念的语义，也不要求覆盖全

为什么需要语义标签？

- 隐式语义特征已经可以很好的帮助推荐
- 语义标签做好的难度和资源投入要远大于隐式语义特征

BUT

- 频道，兴趣表达等重要产品功能需要有一个有明确定义，容易理解的文本标签体系
- 语义标签的效果是检查一个公司NLP技术水平的试金石

典型的层次化文本分类算法

元分类器类型：

- SVM
- SVM + CNN
- SVM + CNN + RNN


实体词识别算法
英超-利物浦0-0曼联，德赫亚频频开挂
分词&词性标注
英超 N 利物浦 N 0-0 曼联 N ，德赫亚 N 。。。
抽取候选
英超联赛
利物浦足球俱乐部
利物浦市
曼联俱乐部
德赫亚
。。。
去歧
英超联赛
利物浦足球俱乐部
曼联俱乐部
德赫亚
。。。
计算相关性
新版实体词
展开>>
大卫·德赫亚 0.9973
利物浦足球俱乐部 0.9899
曼彻斯特联足球俱乐部 0.9835
英格兰足球超级联赛 0.9565
兹拉坦·伊布拉希莫维奇 0.6718
卢克·肖 0.6559
韦恩·鲁尼 0.6387
埃姆雷·詹 0.6320
保罗·博格巴 0.6196
迈克尔·卡里克 0.5185

头条用户标签概览

- 兴趣特征：
  - 感兴趣的类别和主题
  - 感兴趣的关键词
  - 感兴趣的来源
  - 基于兴趣的用户聚类
  - 各种垂直兴趣特征（车型，体育球队，感兴趣股票）
- 身份特征：
  - 性别
  - 年龄
  - 常驻地点
- 行为特征：
  - 晚上才看视频

主要有哪些策略?

- 过滤噪声：过滤停留时间短的点击，打击标题党
- 惩罚热点：用户在热门文章上的动作做降权处理
- 时间衰减：随着用户动作的增加，老的特征权重会随时间衰减，新动作贡献的特征权重会更大
- 惩罚展现：如果一篇推荐给用户的文章没有被点击，相关特征（类别，关键词，来源）权重会被惩罚
- 考虑全局背景：考虑给定特征的人均点击比例（做L1norm）

用户标签批量计算框架

- 每天抽取昨天使用过头条的用户
- 抽取这些用户过去两个月的动作数据
- 在Hadoop集群上批量计算结果

批量计算用户标签的问题

计算量太大！随着：

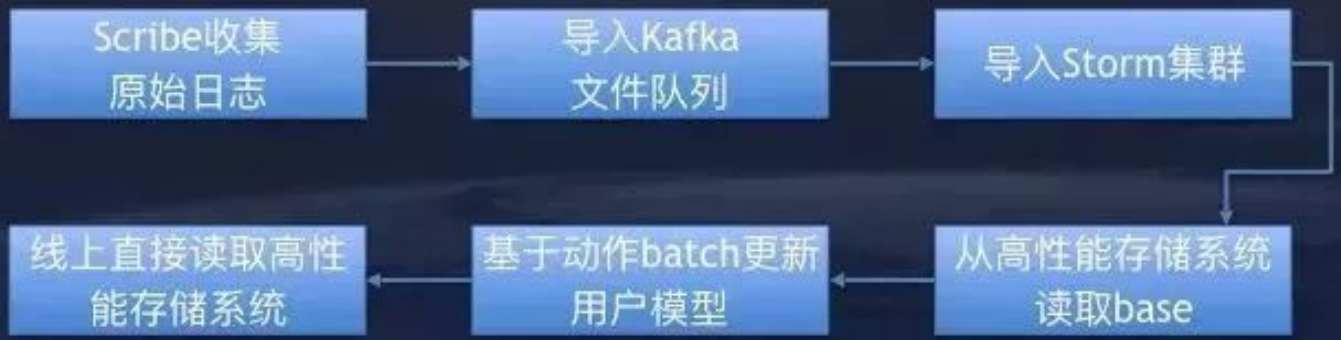
- 用户的增长
- 兴趣模型种类的增加
- 其它批量处理任务的增加

导致：

- 批量处理任务当天完成的越来越勉强
- 集群计算资源紧张影响其它工作
- 集中写入分布式存储系统的开销越来越高
- 用户兴趣标签更新延迟越来越高

用户标签流式计算框架

- 用Storm集群实时处理用户动作数据
- 每收集一定量（batch）的用户数据就重新计算一次用户兴趣模型
- 用大规模+高性能存储系统支持用户兴趣模型读写



流式计算和批量计算混合使用

- 大部分user profile采用流式计算
  - 各个粒度的兴趣标签
  - 垂直领域profile
- 对时效性不敏感的user profile采用Batch计算
  - 性别，年龄
  - 常驻地点

对推荐效果可能产生影响的因素

- 候选内容集合的变化
- 召回模块的改进和增加
- 推荐特征的增加
- 推荐系统架构的改进
- 算法参数的优化
- 规则策略的改变

我们需要：

- 完备的评估体系
- 强大的实验平台
- 易用的实验分析工具

推荐评估体系需要注意的问题

兼顾短期指标和长期指标

兼顾用户指标和生态指标

注意协同效应的影响，有时候需要做彻底的统计隔离

为什么需要一个强大的实验平台

同时在线的实验多：每天数百个

高效管理和分配实验流量

降低实验，分析成本，提高算法迭代效率


A/B Test实验系统原理
流量分桶
分配实验流量
分配实验组

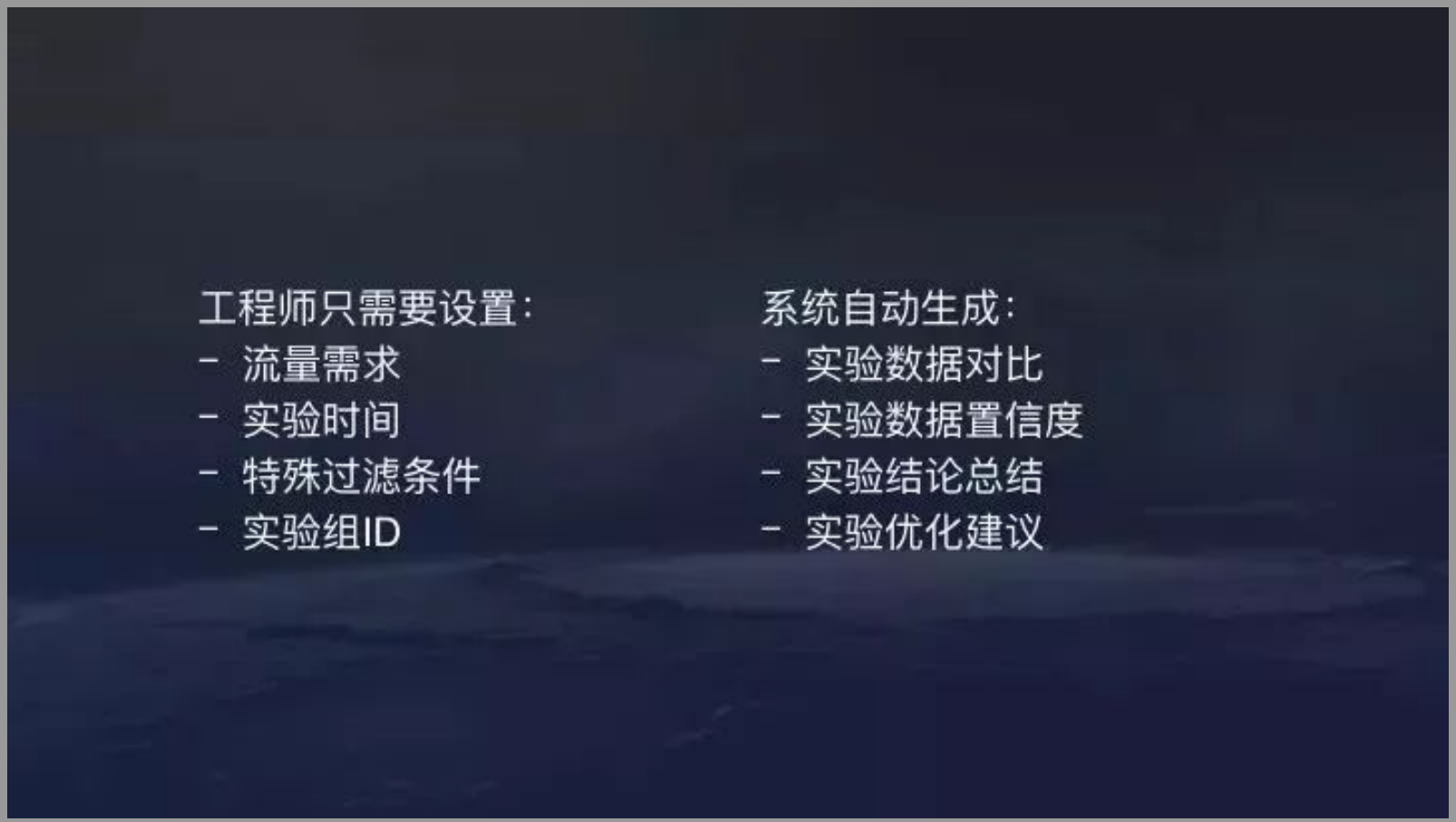
实验数据统计分析

动作收集

日志处理

分布式统计

写入数据库


工程师只需要设置：
- 流量需求
- 实验时间
- 特殊过滤条件
- 实验组ID
系统自动生成：
- 实验数据对比
- 实验数据置信度
- 实验结论总结
- 实验优化建议

人工抽样评估分析

线上实验平台只能通过指标变化推测用户体验

数据指标和用户体验存在差异

重大改进需要人工评估二次确认

头条利用内部和外包团队进行例行的人工抽样评估

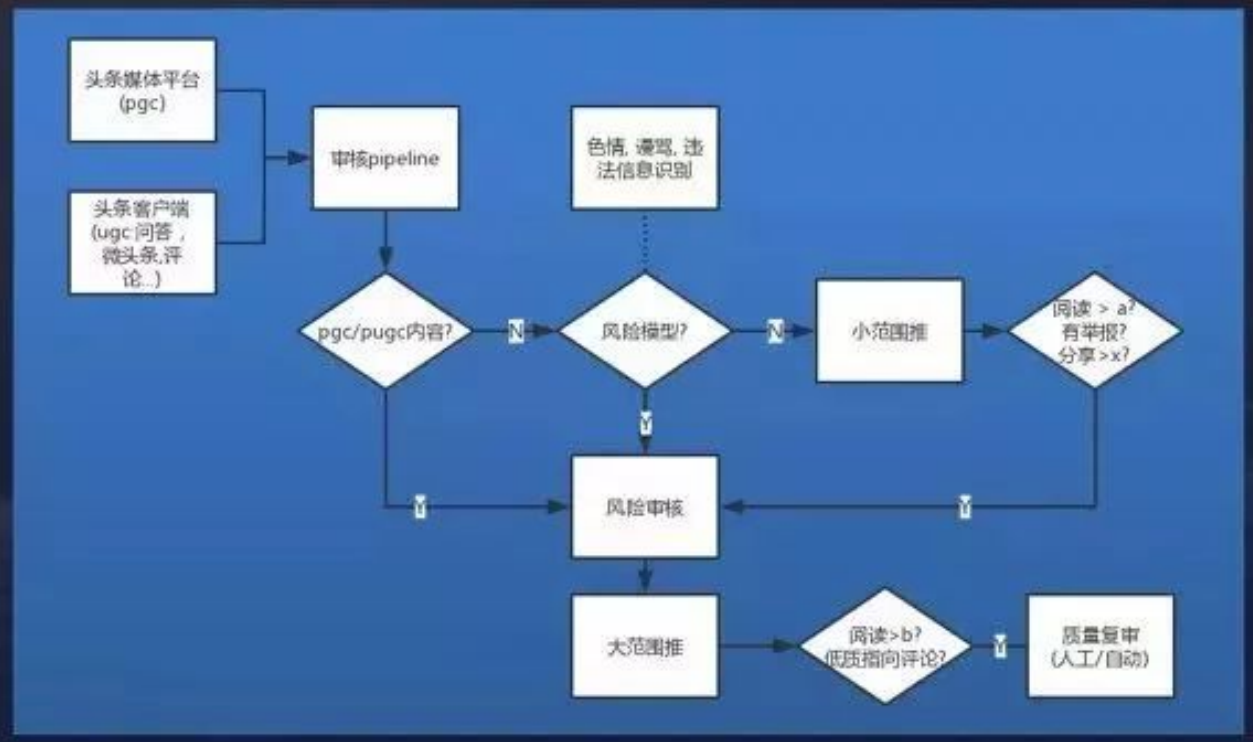
头条的社会责任

今日头条已经成为国内最大的综合资讯平台

如果1%的推荐内容出现问题，就会产生较大的社会影响

头条从创立伊始就把内容安全放在公司最高优先级队列

头条的内容安全机制


头条媒体平台(pgc)
头条客户端(ugc:问答，微头条,评论...)
审核pipeline
色情, 谩骂, 违法信息识别
pgc/pugc内容?
N
风险模型?
N
小范围推
阅读 > a?
有举报?
分享>x?
Y
Y
Y
风险审核
大范围推
阅读>b?
低质指向评论?
Y
质量复审(人工/自动)

风险内容识别技术

- 鉴黄模型：构建了千万张图片样本集，通过深度学习算法(ResNet)训练，召回率99%。
- 低俗模型：对文本和图片同时分析，样本库超过百万，准确率80%+，召回率90%+。不仅处理文章，也对评论做低俗识别。
- 谩骂模型：净化产品评论氛围，识别出不当评论，样本库超过百万，召回率95%+，准确率80%+。

泛低质内容识别技术

- 低质模型是通过对评论做情感分析，结合用户其它的负反馈信息（举报、不感兴趣、踩）等信息，来解决很多语义上的低质问题，诸如题文不符、有头无尾、拼凑编造、黑稿谣言等。
- 目前低质模型的准确率为70%，召回率为60%，结合人工复审召回率能做到95%。

流量池分配

抖音采用的是流量池分配技术，如果你的作品比较好，那么抖音会推送越来越多的流量给你的视频。这就是为什么我们刚发布的视频，只是视频不违规，就会有几百个播放量。

层层推进机制

当你的作品进入流量池后，接下来就看你的作品的表现，如果你的作品表现比较好，那么抖音会将你的作品推动到几千的流量池中，给你分配几千的流量到你的视频。如果你的作品表现不好，那就降低流量的推荐。

视频评价

- 爱心数量：视频右侧有个爱心，点击一下（或者双击屏幕），就会收藏到自己的喜欢列表，爱心越多，越有利于热门。分析为什么会双击，乐意双击，把这个分析清楚，你才能做更优质作品
- 观看时长：为什么用户会乐意看完，因为你的视频对用户有用。所以你制作的视频必须站在用户的角度思考，制作用户喜欢的视频
- 评论数量：大家为什么评论，多看同行视频的评论，你就能找到如何引导用户评论的方法，评论越多越好，不要随便删除评论。如果影响你推荐，负面的，广告类的评论你可要删除
- 转发量：好的作品，大家才喜欢转发
- 关注：为什么浏览者要关注你，无非就是想下次容易找到你
- 过往权重表现：该账号以前是否有违规，是否被限流