视频推荐搜索中的用户兴趣

优酷 搜索、推荐、内容智能负责人 数据智能部总监 李玉

主办方 Geekbang 极客邦科技 InfoQ

AiCon

全球人工智能与机器学习技术大会

助力人工智能落地

2018.1.13 - 1.14　北京国际会议中心

扫描关注大会官网

极客时间

重拾极客精神·提升技术认知

下载极客时间App

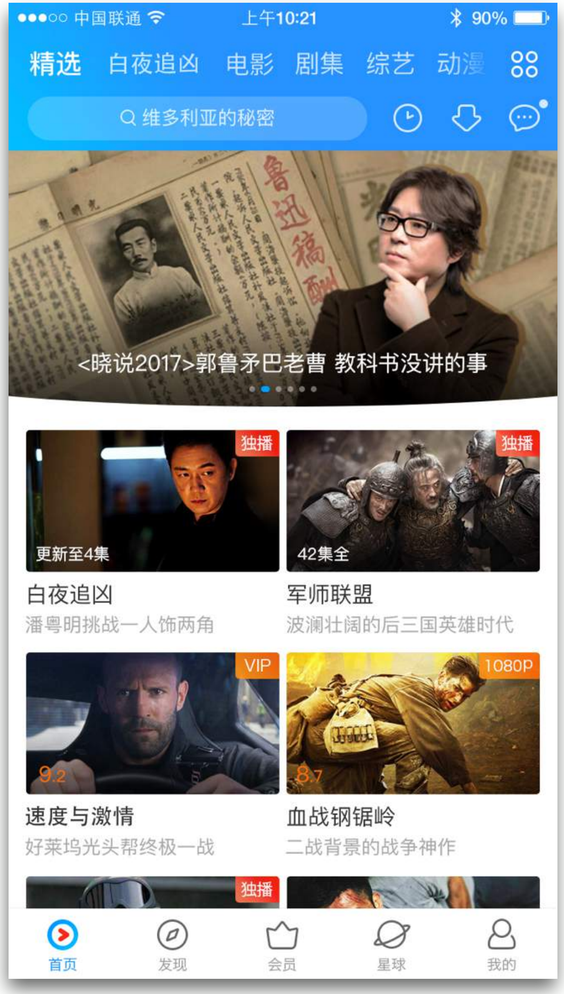
获取有声IT新闻、技术产品专栏，每日更新

扫一扫下载极客时间App

Agenda

- 优酷视频个性化搜索推荐简介
- 视频个性化搜索推荐中的用户兴趣表达的挑战
- 当前工业界常见方法的问题探讨
- 我们的尝试的方法

优酷个性化服务简介


中国联通
上午10:21
90%
精选
白夜追凶
电影
剧集
综艺
动漫
维多利亚的秘密
<晓说2017>郭鲁矛巴老曹 教科书没讲的事
独播
更新至4集
白夜追凶
潘粤明挑战一人饰两角
独播
42集全
军师联盟
波澜壮阔的后三国英雄时代
VIP
9.2
速度与激情
好莱坞光头帮终极一战
1080P
8.7
血战钢锯岭
二战背景的战争神作
独播
首页
发现
会员
星球
我的

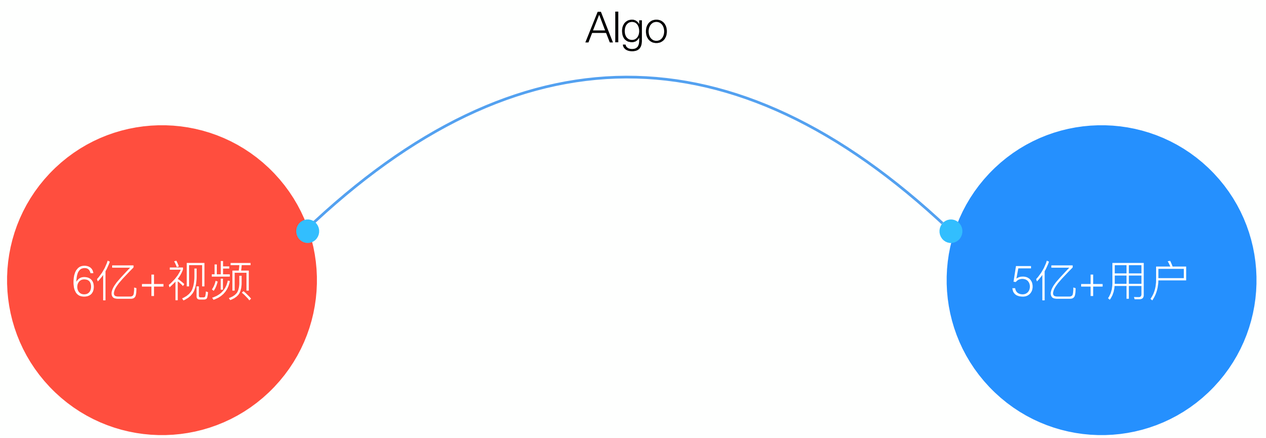

中国联通
上午10:21
90%
精选
白夜追凶
电影
剧集
综艺
动漫
爸爸去哪儿
<小狗牵小手>道格苏苏超有爱
32集全
我的宝贝
佟大为幸福的奶爸生活
34集全
三个奶爸
迥异的三人爆笑带娃之路
更新至20170826期
变形计
青少年的生活角色互换
全13期
爸爸去哪儿第四季
温暖的亲子互动真人秀
独播
独播
首页
发现
会员
星球
我的


今日推荐


推荐
搞笑
音乐
影视剧
动漫
社会
截获了十几亿中国人的信息,奥巴马最后一句亮了
希拉里重返讲台,惊斥特朗普坑害美国必被弹劾下台

个性化服务在优酷

Data…



- 一多半的视频播放通过个性化搜索推荐技术分发
- 对于CTR、人均播放量、人均时长、留存率等均有显著提升
- 帮助用户发现好内容，帮助高质量内容触达精准受众

视频推荐中用户兴趣表达的挑战

视频推荐的用户兴趣表达的挑战

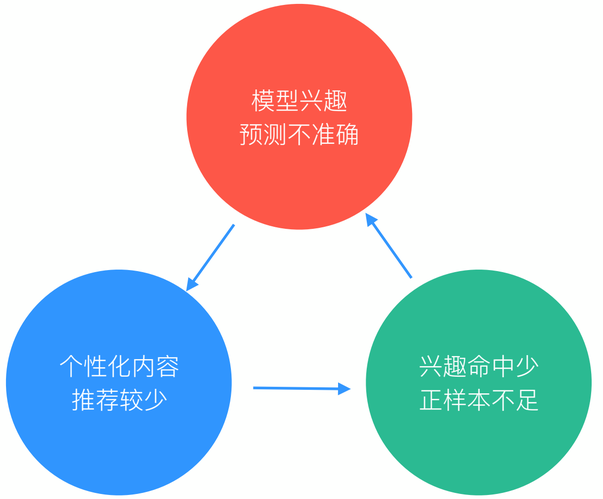
- **技术挑战：**
  - **剧、综、影、漫：**用户选择成本高，用户追的剧、综艺少，推荐成功率低
  - 用户目的性强，发现、浏览、逛的心智低
  - 长节目可选择空间有限
  - 头部节目用户行为稀疏，大量用户每月只观看3个以下节目，**对比**：
    - **短视频信息流场景：**通过数百个观看行为推荐30个
    - **优酷头部节目：**通过3、4个观看行为推荐30个
  - 数据噪声多、分布驱热、highly biased，常用推荐算法模型描述能力不足

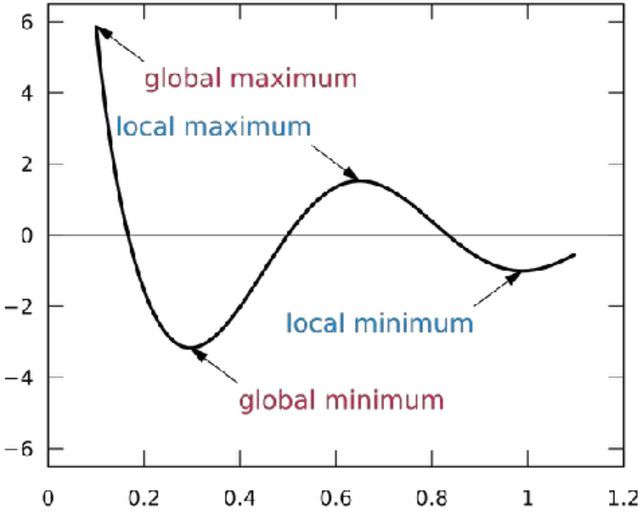
视频推荐的用户兴趣表达的挑战 cont.

- **技术挑战：**
  - 视频内容兴趣复杂，感性、微妙、亚文化细分多样，对于符合兴趣大方向的惊喜度（serendipity）与多样性要求更高，**对比：**
    - **电商：**兴趣明确：想买4K电视、牛仔裤、连衣裙；高度结构化，类目体系清晰
    - **视频：**
      - 兴趣感性、微妙：喜欢香港武侠片但是讨厌成龙; 喜欢日本动漫，今敏等、但讨厌宫崎骏；
      - 兴趣会进化、发展、细分，如：
        - 相声：郭德纲 小岳岳-》方清平；或者-》王玥波评书；或者-》侯宝林 刘宝瑞 马三立 传统
        - 科幻迷：从浅度：看星战、地心引力-》中度：星际穿越-》深度：银翼杀手、降临、三体；
      - 微妙的亚文化：二次元、游戏、直播；文艺青年；腐、柜；追剧族、韩剧迷、恐怖片迷
      - 兴趣体现的是用户的个人认同
      - 兴趣多维度正交，如：
        - 只看”大制作”、美剧质感
      - 不喜欢重复，期待惊喜（serendipity）

识别、表达用户兴趣的重要性

- Retargeting（看了又看）：
  - 推荐用户有过交互的内容（看了又看）
  - 成功率高，长期价值低
    - 局部提升非全局提升（抢其他渠道流量）
  - 成功率高因此ctr高
  - 容易陷入局部最优
- 热点推荐
  - 推荐近期热点
  - 容易陷入局部最优
- 个性化兴趣推荐
  - 推荐符合每个用户兴趣的内容
  - 成功率低因此ctr偏低
  - 更具长期价值
  - 短期收益可能小，但容易长期收敛
- **推荐命中成功率：**retargeting > 热点 > 个性化发现
- **推荐命中（不命中）价值：**个性化发现 > 推荐热点 > retargeting





当前工业界常见方法的问题探讨

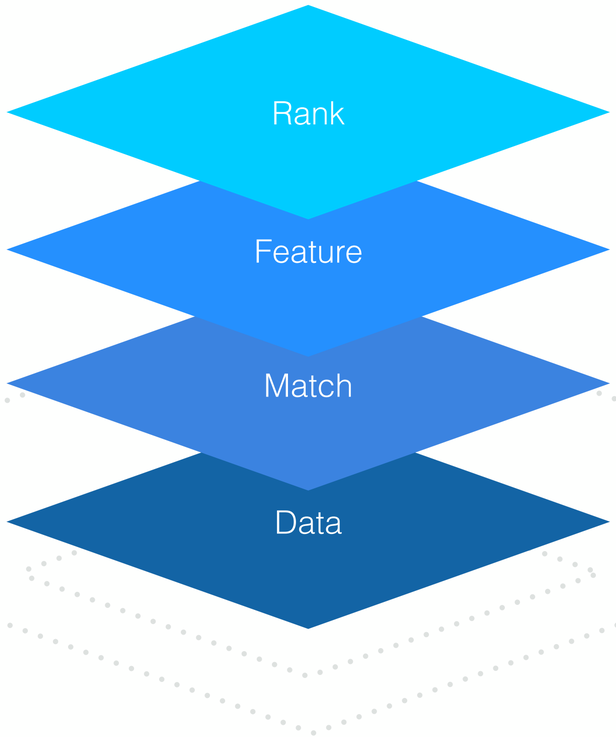
个性化推荐工业界常用方法

- 流程：召回、排序
- 特征：
  - 统计特征
  - 用户画像：DEMO、用户对于标签的frequency、recency
  - 高维组合特征
  - Item based similarity(i2i)

Common Algo Framework(对应的优酷的方法)

Rank R

FTRL, DNN, XGBoost, FFM
Ensemble
Rerank



F Feature

Item/User/User2Item Statistics
User Profile: (Demo,Interest profile,
search profile, view history)
Item tags, categories, topics
item/tag/topic relevance scores

Match M

Item Based CF, DNN CF
Slim CF
Tag to Item,
User2user2Item
Star2Item
Popularity, Trending

D Data

ETL
offline/streaming

常用方法对于表达用户视频兴趣的问题

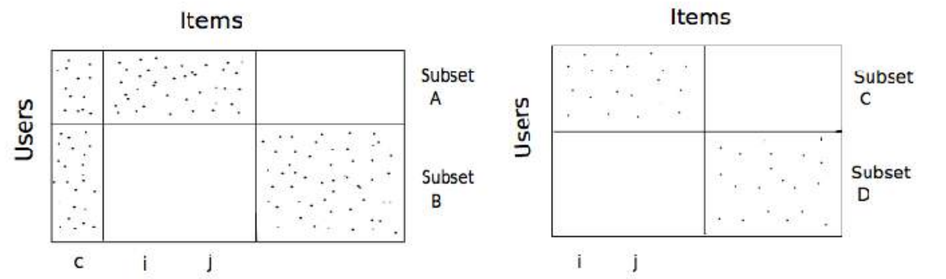
- Demo（年龄、性别、地域），设备类型、城市...
  - 问题：用户的内容兴趣与以上信息相关性不大
  - 问题：三线城市50岁男性可能和一线城市30岁女性的观看习惯一致
- 基于内容标签的用户画像
  - 人工内容标签：恐怖片、动作片、搞笑、香港片、韩国片
  - Topic Modeling标签：LDA提取视频标题、描述的主题（内容数据噪声大）
  - 基于统计的方法（frequency、recency）建立用户标签
  - 问题：人工标签主观性大、噪声大
  - 问题：人工标签粒度容易过于宽泛
  - 问题：topic modeling标签噪声大、数据稀疏
  - 问题：往往基于统计的方法，很难精准描述用户的兴趣
  - 问题：容易受到驱热的影响

常用方法对于表达用户兴趣的问题 cont.

- 高维组合特征
  - 通过组合以上各种特征，产生更丰富的信息
  - 问题：容易受到噪声影响
  - 问题：计算量过大
- Item based similarity (i2i)
  - CF similarity
  - SVD++/MF
  - Slim
  - DNN
  - 简单高效

Problem of I2I

- Item based CF是学术和工业界都最有效的方法之一
- Item based方法比User based方法更有效。
  - 主要因为user 维度行为更稀疏，噪声更大。Item的维度积累历史行为更多，variance更小。
- **问题1:** 由于基于item维度的全局统计，每个用户观看item的不同原因信息被平均掉。对于一个视频，有的用户因为热度观看，有的用户因为主题的类型观看，有的用户因为主演、导演观看。
- **问题2:** 不同用户群体的不同喜好在全局Item similarity的计算过程中被平滑掉。
- **问题3:** 对于长尾item行为数据过于稀疏
- **问题4:** 粒度太细，数据稀疏，扩展能力弱
- **问题5**: 驱热、哈利波特现象



(a) Overlapping rated items between user subsets (b) No common rated items between user subsets

介绍我们的一些尝试

基础用户画像做法

用户观看行为

用户对于各类标签观看的
Frequency / Recency

内容

内容的标签、类目体系
演员、导演等Metadata

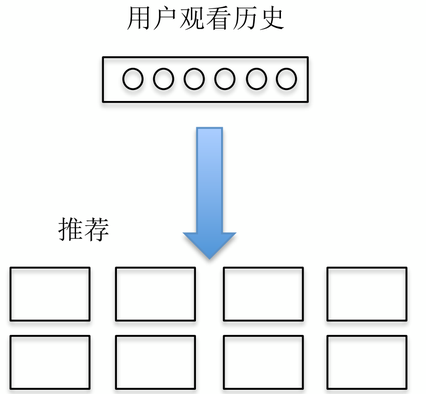
兴趣画像

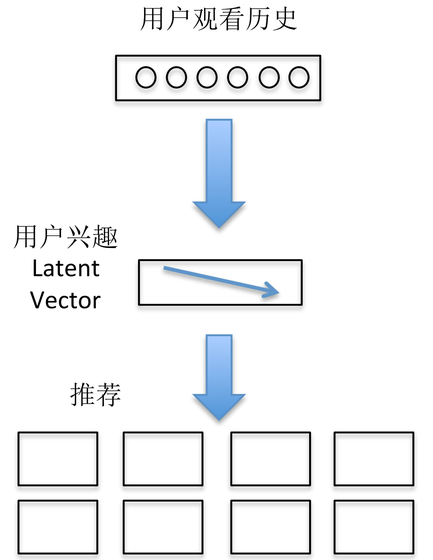
针对每个标签、类目的兴趣强度
分

问题： 基于统计，无法区分驱热、类型、明星等信息
粒度过于粗

User Interest Latent Vector

- End2End 黑盒模型由于噪声与概率分布假设的问题并非全局收敛，需缩小搜索空间
  - 拆解为多个更容易的子问题
  - 机器学习解一个End2End大问题 < 拆解为若干个更容易的小问题
- 传统End2End方法易受数据稀疏与噪声影响：
  - End2End模型：观看历史<->节目推荐，易受噪声影响
  - 拆解为子问题预测模型：
    - 观看历史<->宽泛兴趣分类Latent Vector<->节目推荐，对于噪声更鲁邦
- 宽泛兴趣Latent vector——人工构建类目体系+审核，降噪





用户兴趣的建模的work - CTR

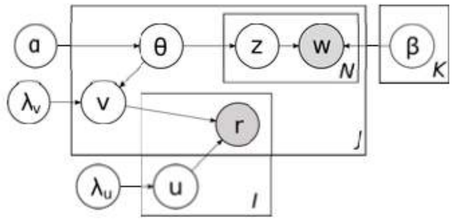
- Collaborative Topic Modeling for Recommending Scientific Articles

1. For each user i, draw user latent vector $u_i \sim \mathcal{N}(0, \lambda_u^{-1} I_K)$.

2. For each item j,
   (a) Draw topic proportions $\theta_j \sim \text{Dirichlet}(\alpha)$.
   (b) Draw item latent offset $\epsilon_j \sim \mathcal{N}(0, \lambda_v^{-1} I_K)$ and set the item latent vector as $v_j = \epsilon_j + \theta_j$.
   (c) For each word w_{jn},
      i. Draw topic assignment $z_{jn} \sim \text{Mult}(\theta)$.
      ii. Draw word $w_{jn} \sim \text{Mult}(\beta_{z_{jn}})$.

3. For each user-item pair (i, j), draw the rating

$$r_{ij} \sim \mathcal{N}(u_i^T v_j, c_{ij}^{-1}). \quad (6)$$



用户兴趣的建模的work - CTPF

- Content-based recommendations with Poisson factorization

1. **Document model:**
   (a) Draw topics $\beta_{vk} \sim \text{Gamma}(a, b)$
   (b) Draw document topic intensities $\theta_{dk} \sim \text{Gamma}(c, d)$
   (c) Draw word count $w_{dv} \sim \text{Poisson}(\theta_d^T \beta_v)$.
2. **Recommendation model:**
   (a) Draw user preferences $\eta_{uk} \sim \text{Gamma}(e, f)$
   (b) Draw document topic offsets $\epsilon_{dk} \sim \text{Gamma}(g, h)$
   (c) Draw $r_{ud} \sim \text{Poisson}(\eta_u^T(\theta_d + \epsilon_d))$.

- A Practical Algorithm for Solving the Incoherence Problem of Topic Models In Industrial Applications

用户兴趣的建模的work - CTPF with popularity, stars tags and queries

- 实现性能优化，scalable to internet scale
- 基于parameter server架构的分布式实现
- EM不是全局收敛。针对每个topic进行人工审核，再作为初始值进行迭代。
- 扩展到文本+标签+meta+流行度
- 基于兴趣向量的个性化I2I similarity

CTPF with popularity, stars, tags and search queries

1. **Document model:**
   - (a) Draw topics $\beta_{vk} \sim Gamma(a, b)$
   - (b) Draw document topic intensities $\theta_{dk} \sim Gamma(c, d)$
   - (c) Draw word count $\omega_{dv} \sim Poisson\left(\theta_d^T \beta_v\right)$
2. **Document tag model:**
   - (a) Draw tag topics $\varphi_{vk} \sim Gamma(i, j)$
   - (b) Draw tag count $\tau_{dv} \sim Poisson\left(\theta_d^T \varphi_v\right)$
3. **Document search queries model:**
   - (a) Draw search query topics $\nu_{vk} \sim Gamma(l, m)$
   - (b) Draw search query count $\rho_{dv} \sim Poisson\left(\theta_d^T \nu_v\right)$
4. **Attributes(popularity/stars) model:**
   - (a) Define attributes(popularity and star) set of length s $\{popularity, star_0, star_1, \ldots\}$
   - (b) Draw attributes(popularity and star) topics $\pi_{sk} \sim Gamma(l, m)$
   - (c) Draw attributes(popularity and star) count $\alpha_{dv} \sim Poisson\left(\theta_d^T \pi_s\right)$
5. **Recommendation model:**
   - (a) Draw user preferences $\eta_{uk} \sim Gamma(e, f)$
   - (b) Draw document topic offsets $\epsilon_{dk} \sim Gamma(g, h)$
   - (c) Draw $r_{ud} \sim Poisson\left(\eta_u^T(\theta_d + \epsilon_d)\right)$

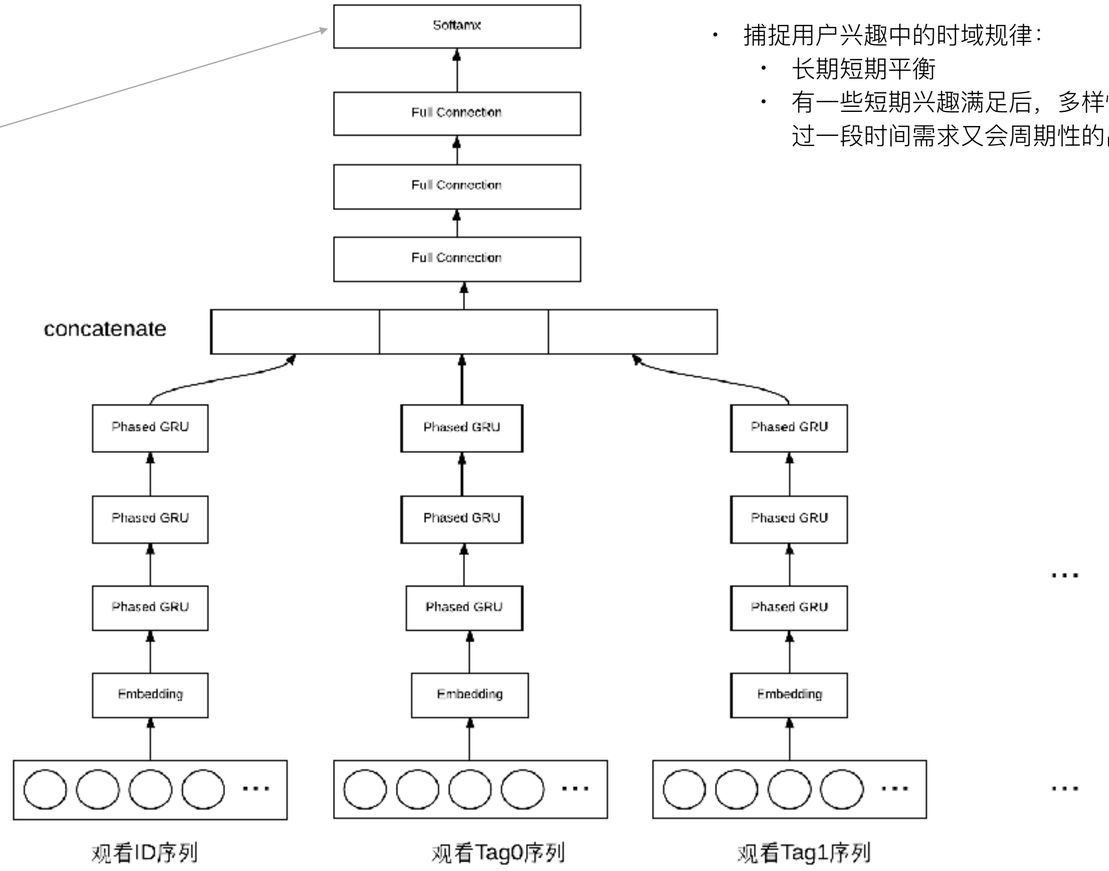
```
Topic488 value:3.307159
--------------------------
喜剧 24.12363362
东北 23.62279569
低俗 18.76789941
二人转 15.72578497
笑星 14.3499465
搞笑 14.09372048
东北话 12.94013481
开心麻花 10.20048983
客串 8.84240383
热闹 7.77842085

Topic433 value:2.875541
--------------------------
速度与激情 19.19315398
飚车 15.48903876
刺激 14.26553517
大片 13.99698715
震撼 12.6045356
赛车 12.54367218
好莱坞 12.53610959
肌肉男 10.76938512
硬汉 10.73588267
极品飞车 10.71208749

Topic90 value:1.488183
--------------------------
好莱坞 8.84228278
大片 8.33831162
震撼 7.04246583
热血 5.70970491
战斗 5.69004143
军事 5.6835127
特效 4.40256621
主旋律 4.38128264
军旅 4.37505014
科幻 4.35832843
```

长期兴趣与短期兴趣的平衡——Phased GRU RecNet

Listwise Loss:
BPR/TOP1 Loss



- 捕捉用户兴趣中的时域规律：
  - 长期短期平衡
  - 有一些短期兴趣满足后，多样性需求会变强
    过一段时间需求又会周期性的出现

Based on: SESSION-BASED RECOMMENDATIONS WITH RECURRENT NEURAL NETWORKS - ICLR2016

长期兴趣与短期兴趣的平衡——Phased GRU RecNet cont.

- GRU:

$$\mathbf{h_t} = g(W\mathbf{x_t} + U\mathbf{h_{t-1}})$$

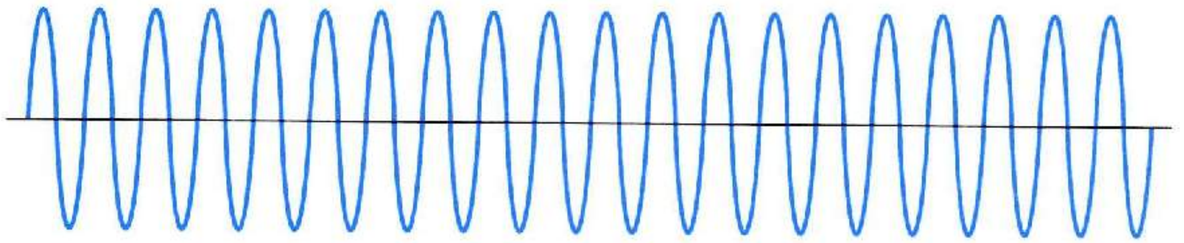
$$\mathbf{h_t} = (1 - \mathbf{z_t})\mathbf{h_{t-1}} + \mathbf{z_t}\hat{\mathbf{h}}_\mathbf{t}$$

$$\mathbf{z_t} = \sigma(W_z\mathbf{x_t} + U_z\mathbf{h_{t-1}}) \qquad \text{update gate}$$

$$\hat{\mathbf{h}}_\mathbf{t} = \tanh\left(W\mathbf{x_t} + U(\mathbf{r_t} \odot \mathbf{h_{t-1}})\right)$$

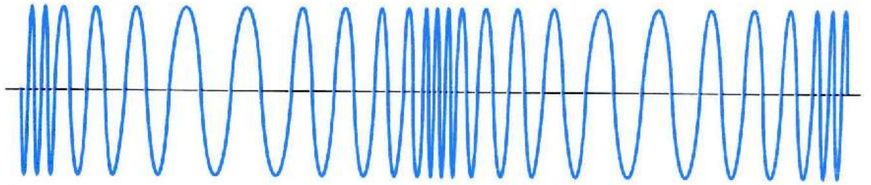
$$\mathbf{r_t} = \sigma(W_r\mathbf{x_t} + U_r\mathbf{h_{t-1}}) \qquad \text{reset gate}$$

- 默认的假设是等距采样:



长期兴趣与短期兴趣的平衡——Phased GRU RecNet cont.

- 用户session实际情况是有的session一天100个行为，有的session一个月只有一个行为

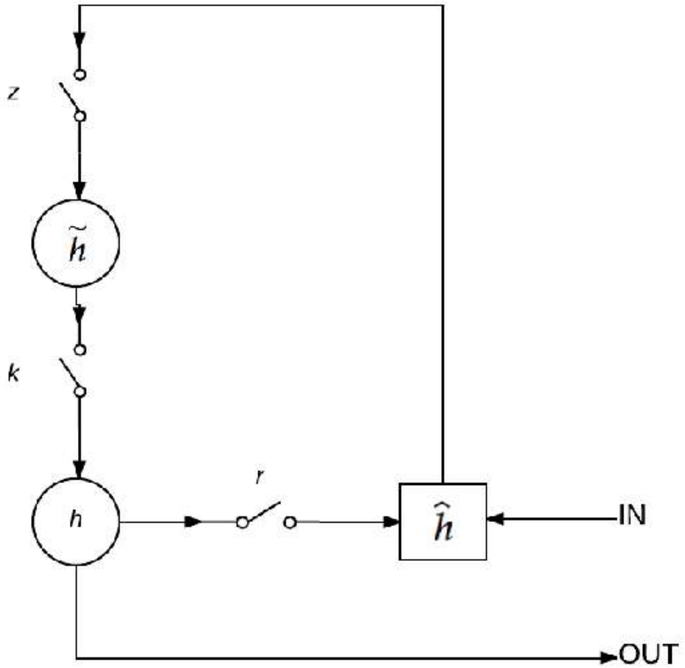


- Phased GRU, 引入time gate k，根据采样间隔控制变量的更新(同时增加一定程度的采样间隔):

$$\phi_t = \frac{(t-s) \bmod \tau}{\tau}, \qquad k_t = \begin{cases} \dfrac{2\phi_t}{r_{on}}, & \text{if } \phi_t < \dfrac{1}{2} r_{on} \\ 2 - \dfrac{2\phi_t}{r_{on}}, & \text{if } \dfrac{1}{2} r_{on} < \phi_t < r_{on} \\ \alpha\phi_t, & \text{otherwise} \end{cases}$$

$$\tilde{h}_j = (1 - z_j) h_{j-1} + z_j \hat{h}_j$$

$$h_t = k_t \odot \tilde{h}_t + (1 - k_t) \odot \tilde{h}_{t-1}$$

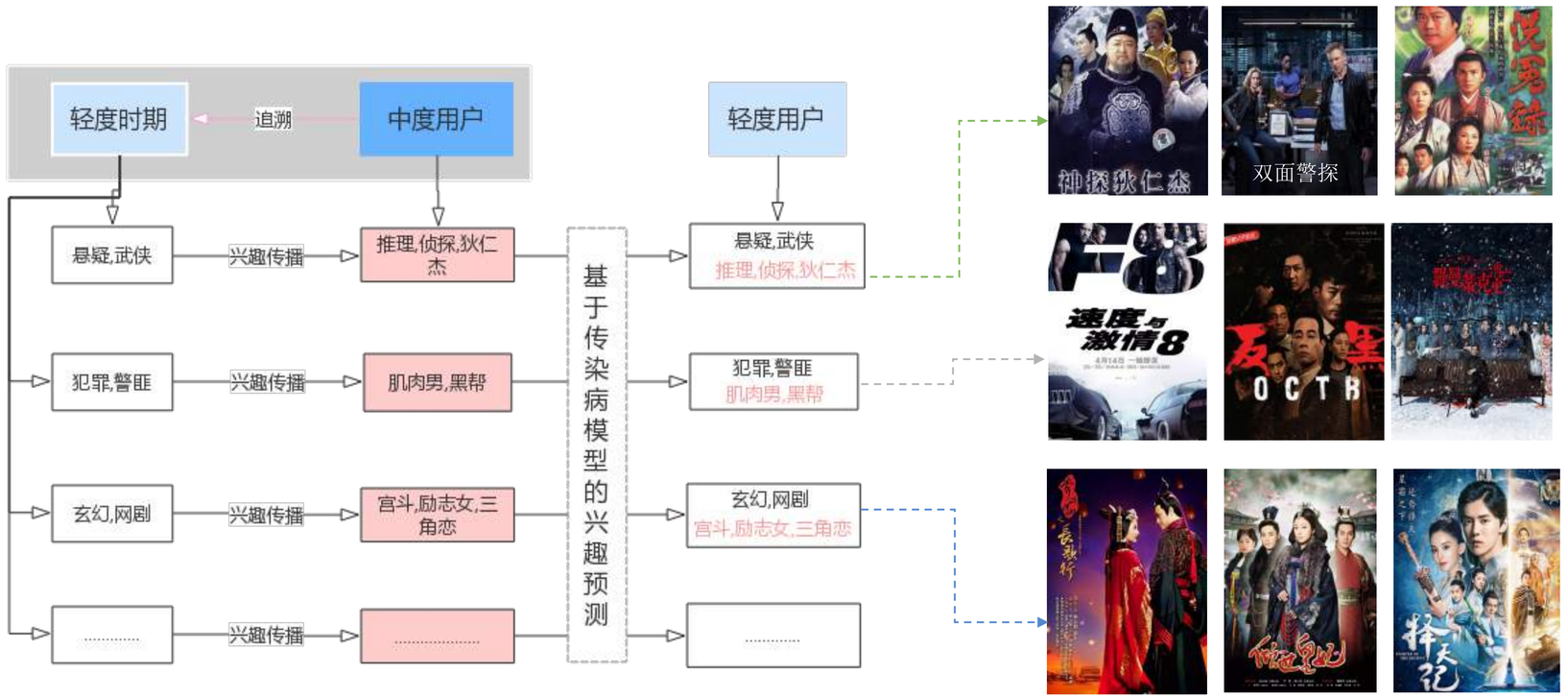


Based on: Phased LSTM: Accelerating Recurrent Network Training for Long or Event-based Sequences

基于传染病模型的有限行为用户兴趣预测

- 大量用户行为非常稀疏，每月观看量不超过3次
- 用户群体的兴趣演变遵循类似传染病传播的机制
- 预测：

$$p(Interest_{t+3\ mon}|Interest_t)$$

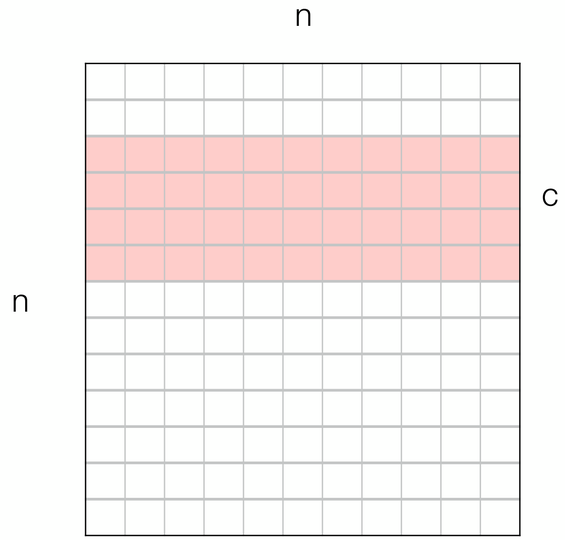


基于Nystrom CUR的exploration

- NxN的I2I矩阵有很多元素很稀疏，explore收集数据需要很多流量，代价很高
- Nystrom CUR:

$$G \in \mathbb{R}^{n \times n} \qquad \tilde{G}_k = C W_k^{+} C^T \qquad C \in \mathbb{R}^{n \times c} \qquad W_k \in \mathbb{R}^{c \times c}$$

- 可以用c个landmark item来代表整个I2I相似度矩阵
- 通过statistical leverage score选择c个item
- 重点explore对于c个item有过观看的用户



基于HIN图、聚类等方法的兴趣识别

- 算法思想

利用用户与节目的播放记录构建二部图，每个节点的标签按相似度传播给相邻节点，在节点传播的每一步，每个节点按照相邻节点的标签来更新自己的标签。与该节点相似度越大，其相邻节点对其标注的影响权值也越大。当绝大多数节点的标签不再更新时，整个网络按照标签就形成了各自所属的社区。

- 权重设定

Item节点的权重为该节目观看人数的倒数
U-I连边的权重为该用户对该节目的观看完成率

User节点的权重为该用户观看节目数量的倒数
U-I连边的权重加入随机因子μ

- 效果评估

将全部用户划分为35830个类簇

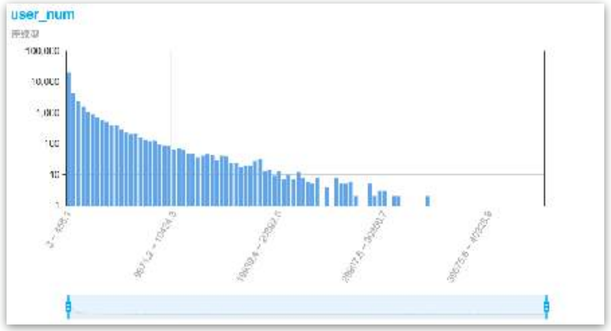
Item在类簇中的挂载成功率为100%

仅有单个Item挂载的类簇占99.48%，最多一个类簇内包含32个节目

类簇内包含的用户个数的分布直方图如右所示，其中最大的类簇包含用户45313个

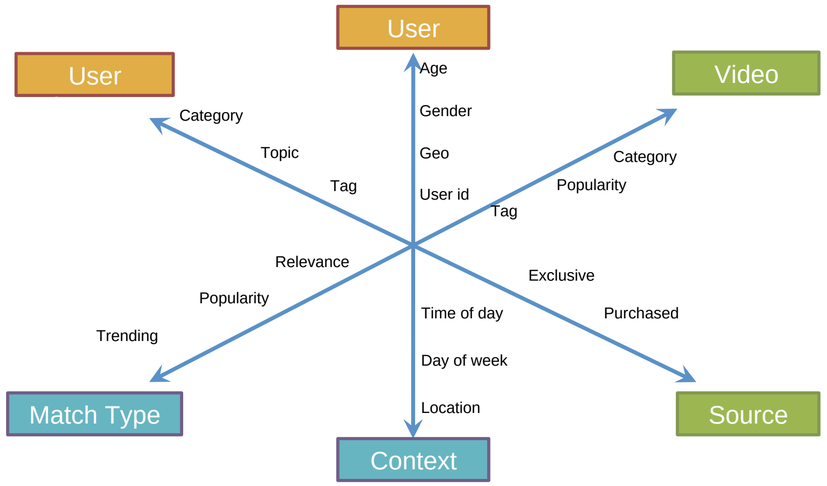
- 典型CASE

序号	节目ID	节目名称
1	323580	汽车城之建筑队
2	323577	汽车城之火车特洛伊
3	318953	和迷你卡车学习
4	323581	汽车城之汤姆的油漆店
5	323573	汽车城之超级变形卡车
6	323571	汽车城之拖车汤姆

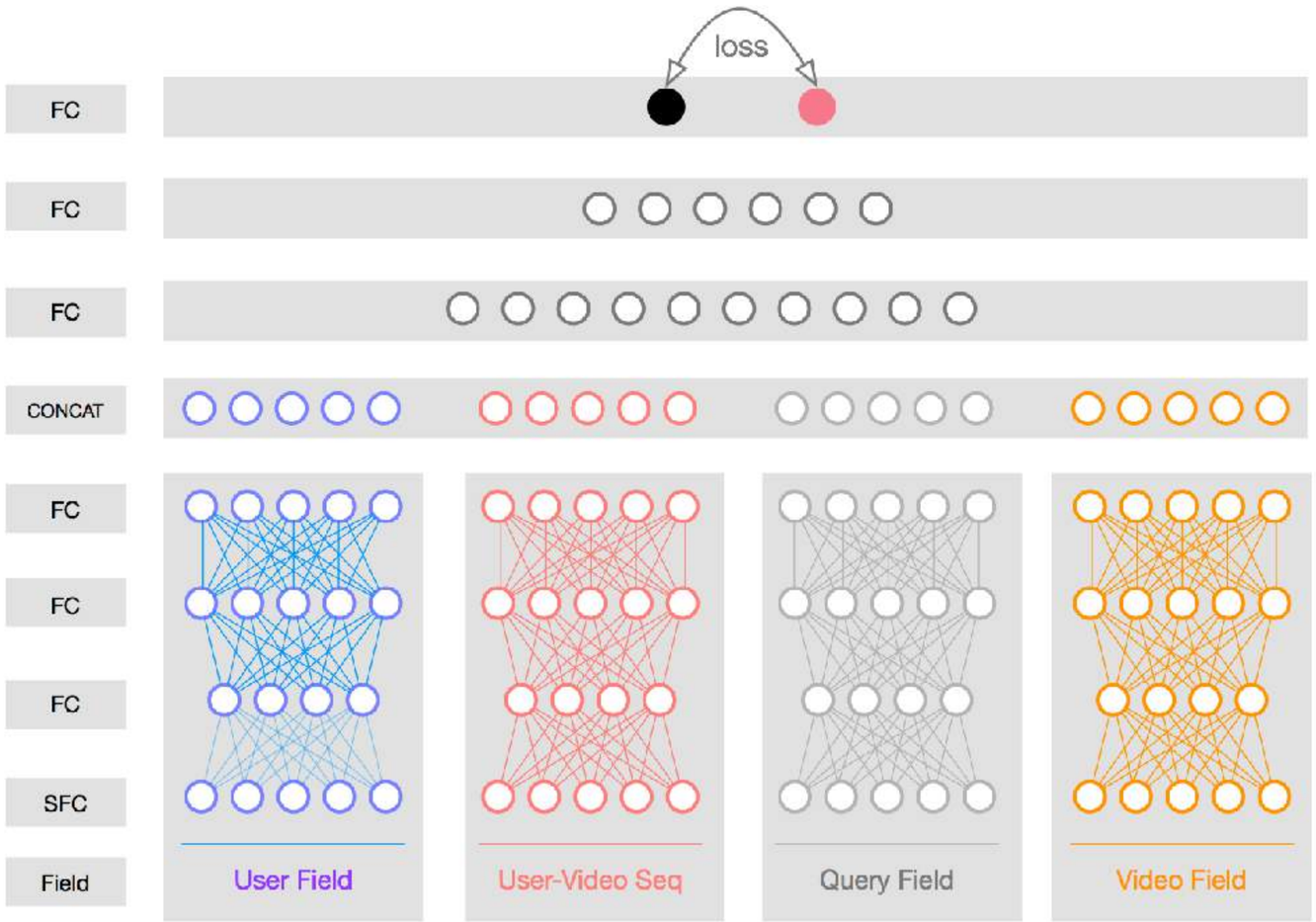


Hierarchical View Feedback Aggregation

- 算法模型能力有限，End2End模型精准capture个性化特征能力有限
- 最优解在非常高纬空间中，由于噪声与模型收敛能力问题，需人工辅助降低搜索空间维度
- 使用交叉特征的统计值，效果好于使用离散交叉裸id特征
- 结合业务理解，辅助模型更好capture个性化特征
- 结合统计量的variance进行噪声过滤
- 交叉统计：更好capture不同用户群体对于不同视频类型的兴趣，如：
  - 爱看韩剧的人群对于台湾偶像剧的人均vv；
  - 爱看日本恐怖片的人群对于美国恐怖片的人均vv；
  - 20岁一线城市女性看游戏人均vv



个性化排序在优酷视频搜索



全局全连接

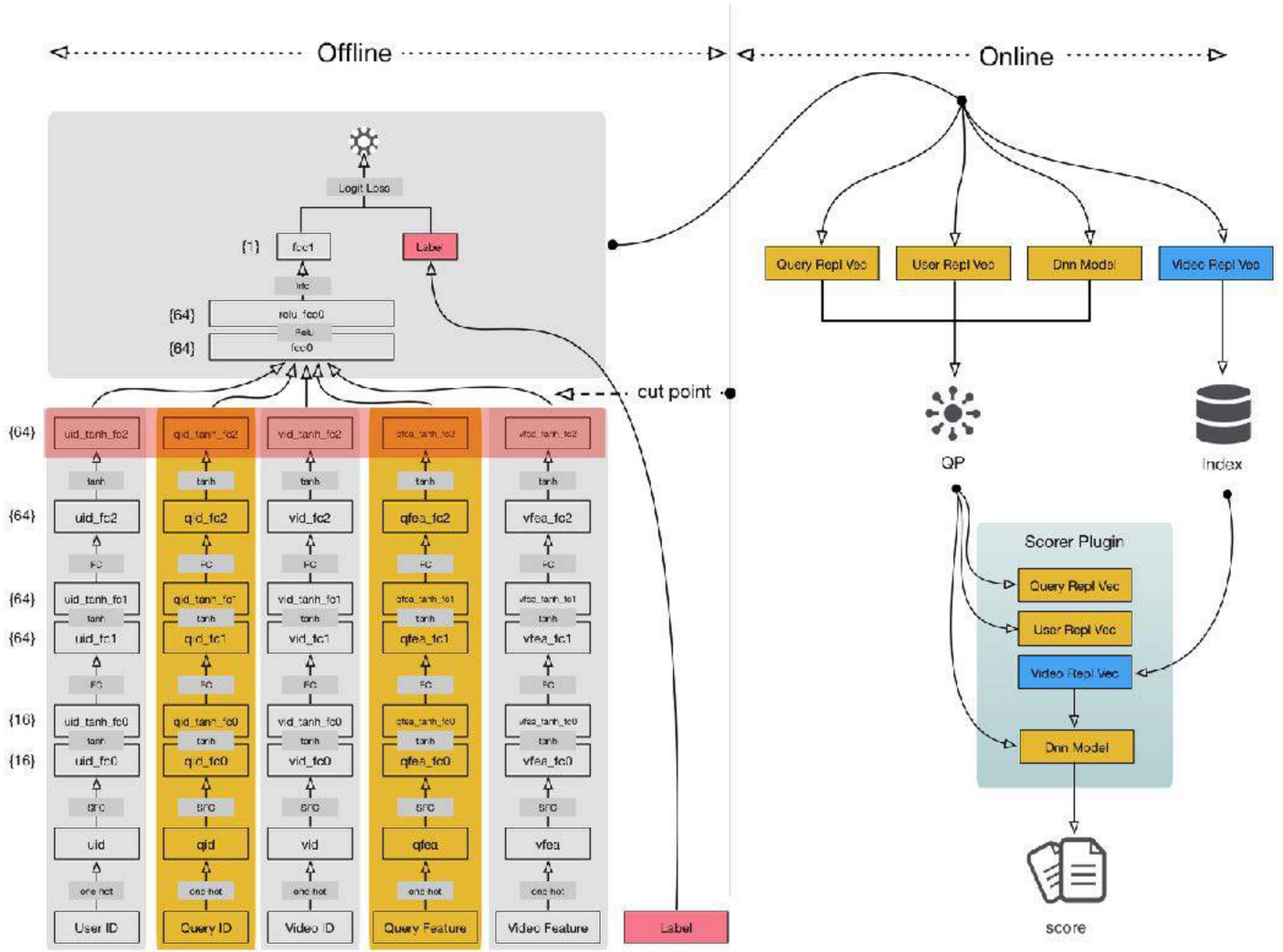
concat

域内信息的二次编码

稀疏全连接

个性化排序在优酷视频搜索-特征域划分及编码

- query user video id域 统计域 用户观看序列 标签兴趣 文本
- 超高维的稀疏编码来表征独立个体
- 利用神经网络来拟合个体共性
- 视频表达是基础
- 按特征的重要度和关联性分域
- 亿级参数
- 挑战：特征维度高 模型存储空间大，离线训练计算时间成本高，在线实现资源占用高，前向网络计算不能满足RT要求
  - 特征分域
  - 随机编码
  - 挂靠编码
  - 抽样技术



这世界很酷

We Are Hiring

ly136216@alibaba-inc.com


Thanks