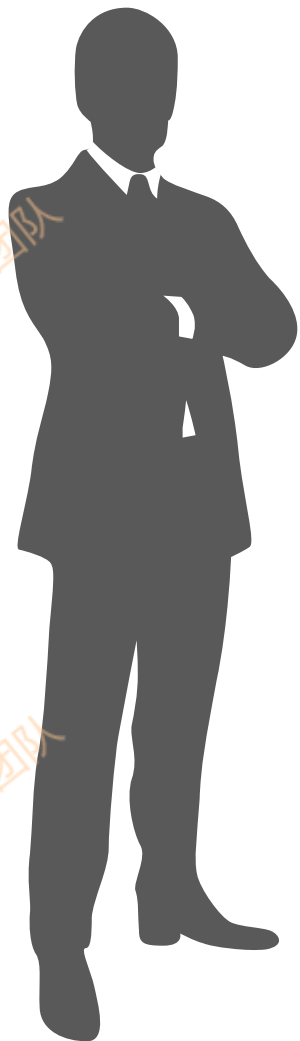


从0到1构建信息流推荐系统

陈辉

内容大纲



一 信息流推荐背景

二 内容理解

三 用户画像

四 召回服务

五 排序服务

六 总结与展望

背景

- feeds流形式
- 海量内容，时效性强
- 由用户触发
- 个性化，千人千面



背景



10条内容

推荐系统

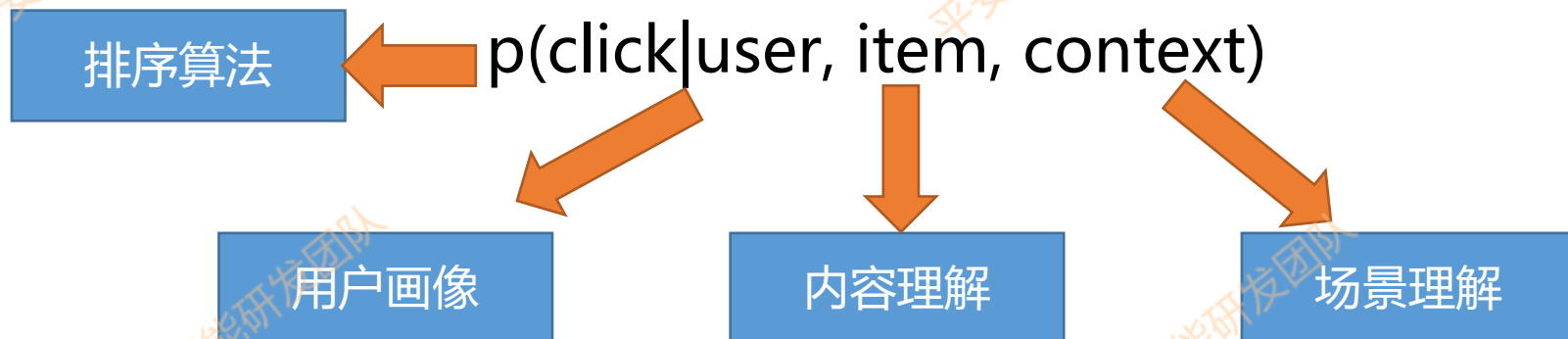
海量内容池(100w+)

- 实时性
- 精准性
- 用户体验

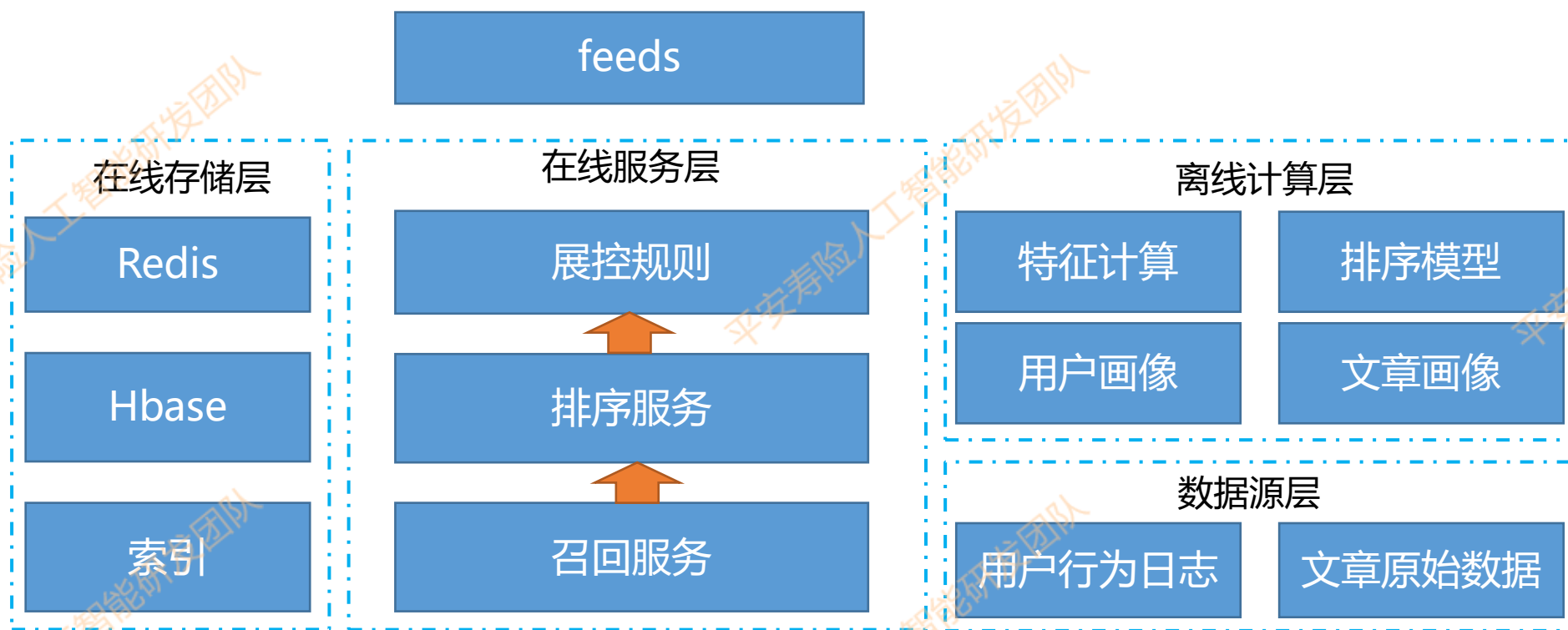
工程与算法的结合

信息流推荐系统基本方法

- 非模型算法：热度、运营...
- 预估用户对内容的CTR，按CTR排序



推荐基本架构



内容理解

- 内容理解是基础
- 基于NLP技术
- 内容分类、关键词...
- topic、embedding

分类：
国际新闻

关键词：
特朗普、伊朗、
普京

特朗普下令袭击伊朗，最后时刻悬崖勒马， 普京发出警告

原创 长安街知事 2019-06-21 18:31:00

随着一架美国无人机被伊朗击落，美伊爆发军事冲突的风险急剧上升。

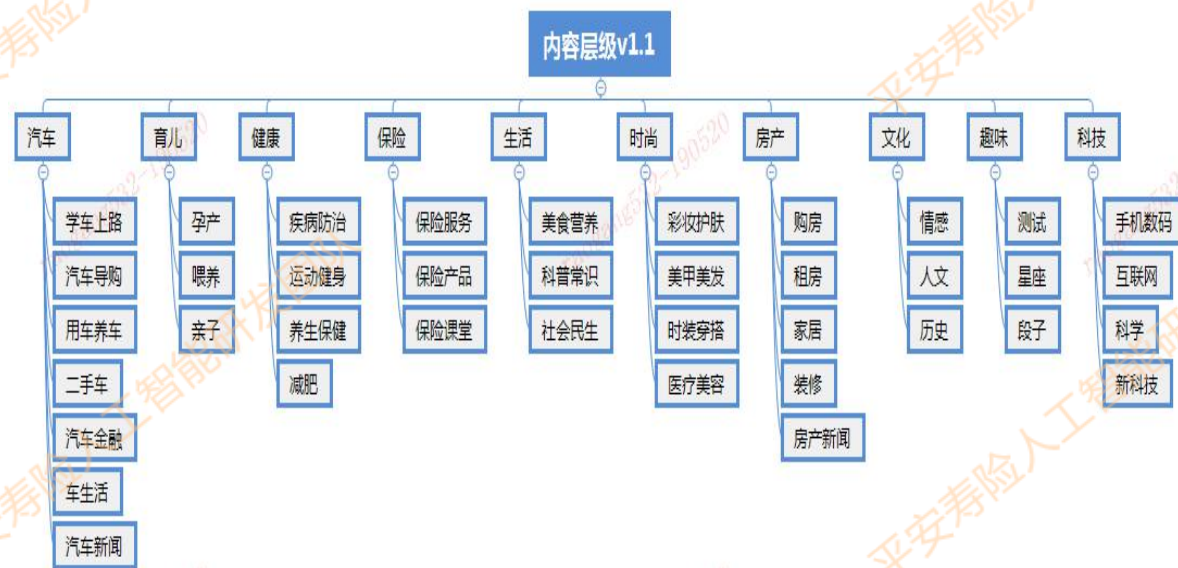
多家美国媒体20日爆料，总统特朗普已经下令对伊朗发动袭击，但就在行动开始前几个小时，特朗普突然改变主意，取消了这一行动。



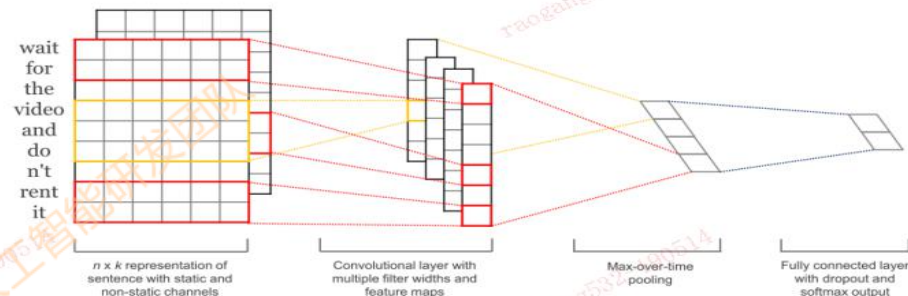
特朗普

内容理解

- 与业务方确定内容层级分类
 - 单标签
- 爬虫获取指定标签内容
- 基于TextCNN训练分类模型
 - 直接对二级分类进行
 - 结果映射到一级分类



分类-TextCNN



用户画像

- 基础画像

- 年龄、性别、地域、学历...

- 兴趣画像

- 长期兴趣
- 短期兴趣

娱乐:0.9, 军事:0.5

平安福:1.0, ...

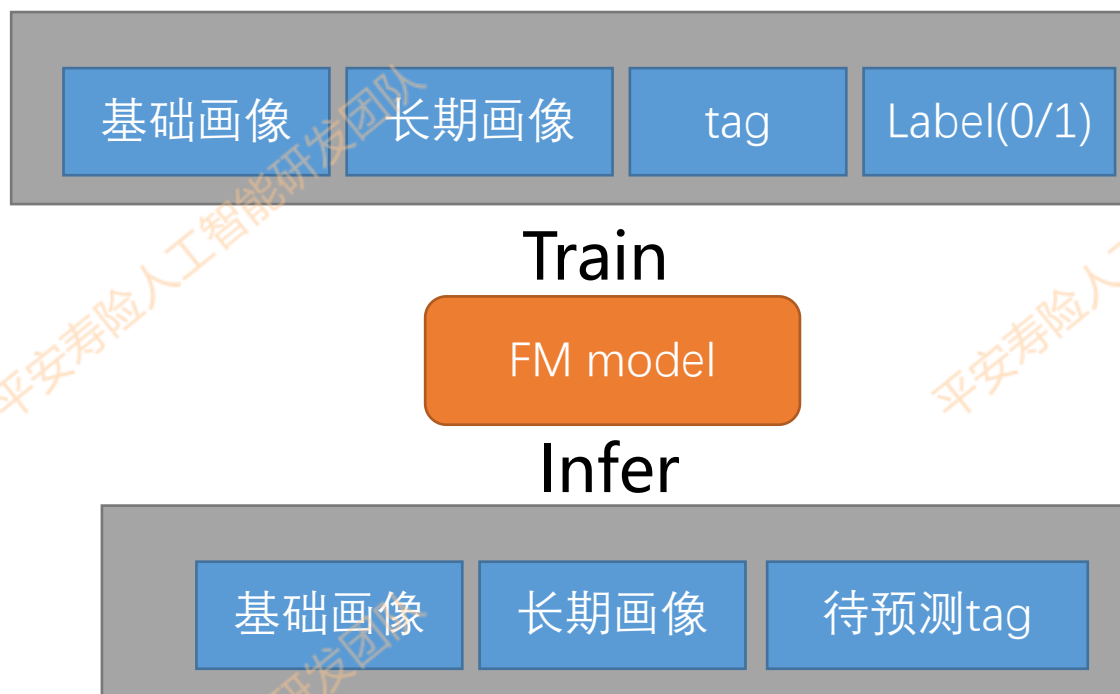
- 兴趣画像计算

- 按天更新，每天累计标签的点击与曝光计数
- 标签点击率作为权重

用户画像

- 探索用户长尾画像

- 曝光不足
- 提升多样性



召回服务

- 高效筛选用户感兴趣内容
 - 100w- > 1k
- 减轻排序服务压力
- 保证多样性

多路召回

- 热点召回和人工运营
- 用户画像召回(Tag base)
- CF召回
- 新内容EE

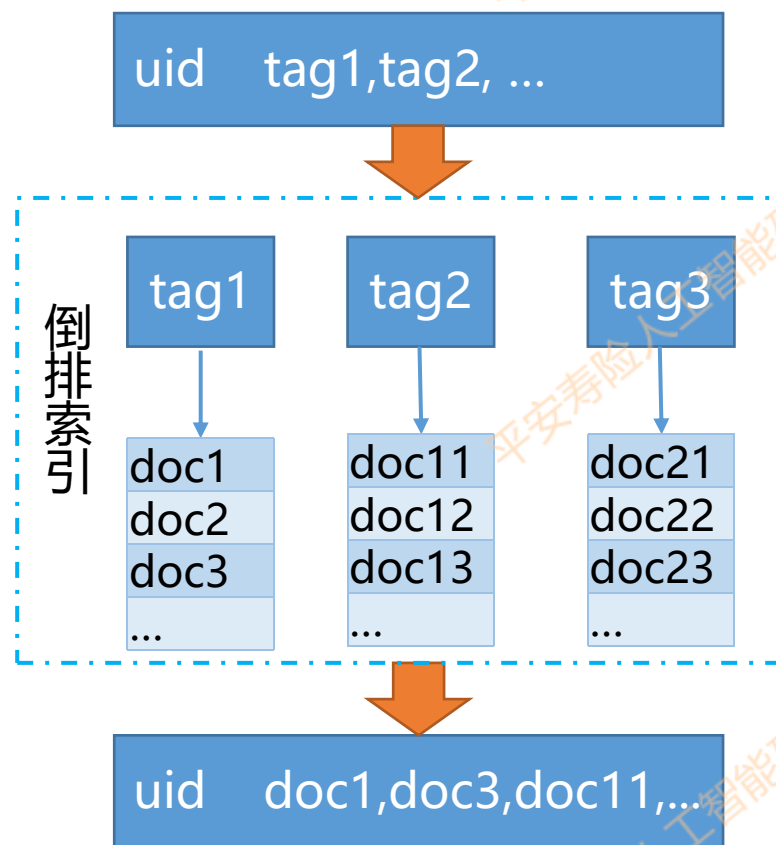
Tag base召回(一)

- 倒排索引
- 倒排分数计算

$$score = w * ctr * \exp(-\Delta t * \beta)$$

w 为文章在tag上的得分， Δt 为文章发布间隔天数， β 为衰减系数

- 每个tag最大召回数 N ，每类tag下最大召回数 M
- 设用户在某tag i 上的归一化权重为 w_i ，则本次召回在该tag下召回的最大文章数为 $n_i = \lfloor w_i * N \rfloor$



Tag base召回(二)

- 截断倒排索引时用“竖切”

- 按每个标签上文章的顺序取文章，如如果有t1, t2, t3个标签，先每个标签取1个，共3个，继续循环每个标签取1个，每步保证： $k_i < n_i$,

$\sum_0^K k_i < M$ 。其中 k_i 为第 i 个标签下召回的文章数



Item CF召回

- 基于行为向量求相似
 - 高维稀疏向量
- 简单可解析性好
- 不能召回新文章
- 结果偏热门

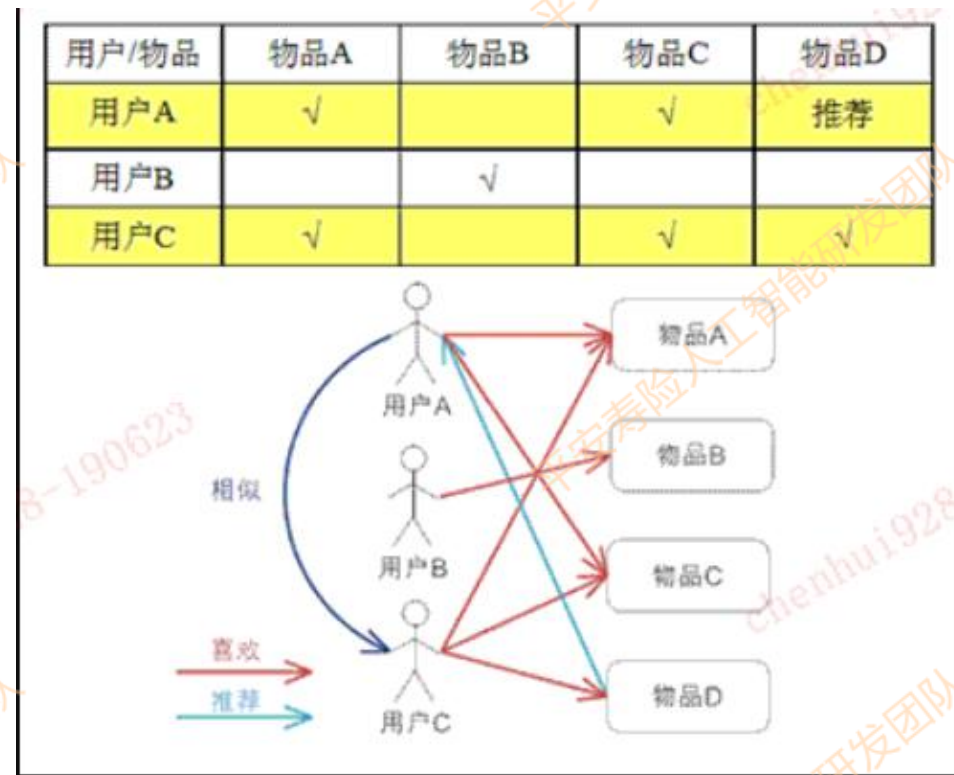
文章j与i的相似度

← W_{ij}

$$W_{ij} = \frac{|N(i) \cap N(j)|}{\sqrt{|N(i)| \cdot |N(j)|}}$$

点击新闻i
的用户集合

同时点击i和
j的用户数

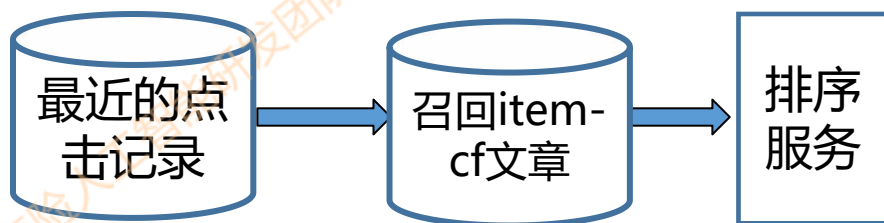


Item CF召回

- 基于内容语义向量求相似
 - 利用NLP处理的结果
 - 对新文章友好
- 大规模稠密向量相似度
 - LSH
- 实时捕捉用户兴趣

文章向量D

$$dis(i, j) = \frac{D_i D_j}{\sqrt{\sum d_i^2 \cdot \sum d_j^2}}$$



User CF召回

- 基于相似用户群体热点
- 实时性强
- 覆盖率高


大规模在线用户CF

- 离线计算每个用户的相似用户top k, 存入cache
- 在线存储每个用户的点击记录
- 在线检索相似用户点击记录

基于点击记录

$$W_{uv} = \frac{|N(u) \cap N(v)|}{\sqrt{|N(u)| \cdot |N(v)|}}$$

$N(u)$ -用户u点击记录


$$W_{uv} = \frac{\sum_{i \in N(u) \cap N(v)} \frac{1}{\log(1 + |N(i)|)}}{\sqrt{|N(u)| \cdot |N(v)|}}$$

惩罚过度活跃用户

新内容EE

- 新文章没有用户行为，无法正常召回
- 需要对新文章进行EE(explore & exploit)
- UCB
- 对每个新item进行实验，然后取ucb-score最大的item进行推荐。关注item回报率（点击），并关注item被探索的次数（没被探索的item有更大的几率被选出）

$$UCB_t(a) = \bar{\mu}_t(a) + \sqrt{\frac{2\log T}{n_t(a)}}$$

等式右边第一部分为item平均回报，点击为1，未点击为0。等式右边第二部分为置信度计算，T为总实验次数，n为a item在t次的时候曝光（实验）次数

排序服务

- 模型：LR, FM, GBDT, DNN...

$$p(\text{click}/\text{user}, \text{item}, \text{context})$$

复杂特征+简单模型

- 线性模型：LR
- 表达能力弱，依赖特征工程
- 训练简单，解析性好
- 上限低

VS

简单特征+复杂模型

- 非线性模型：FM, GBDT, DNN...
- 表达能力强，起点高
- 训练慢，解析性差
- 容易过拟合，难优化，上限高

排序模型

- LR CTR预估(都是离散特征)

$$p(click|x) = 1 / (1 + \exp(w_0 + w_1x_1 + w_2x_2 + \dots))$$

- 特征权重正比特征ctr

$$p(click|x_1) = 1 / (1 + \exp(w_1x_1))$$

- item id作为特征就是使用了该item的ctr特征
- 没有user-item交叉特征，没有个性化
- 交叉特征是提升LR模型效果的关键

FM模型

- FM模型

$$y(x, w, v) = w_0 + \sum_{i=1}^n w_i x_i + \sum_{i=1}^{n-1} \sum_{j=i+1}^n \langle v_i, v_j \rangle x_i x_j$$

- 用隐向量点积表达交叉特征权重
- 无需人工设计交叉特征
- 解决交叉特征稀疏性问题
 - LR : x_i 与 x_j 必须都不为0才能学习两者交叉特征权重
 - FM : v_i 不为0的交叉特征都能作为样本来学习

DeepFM模型

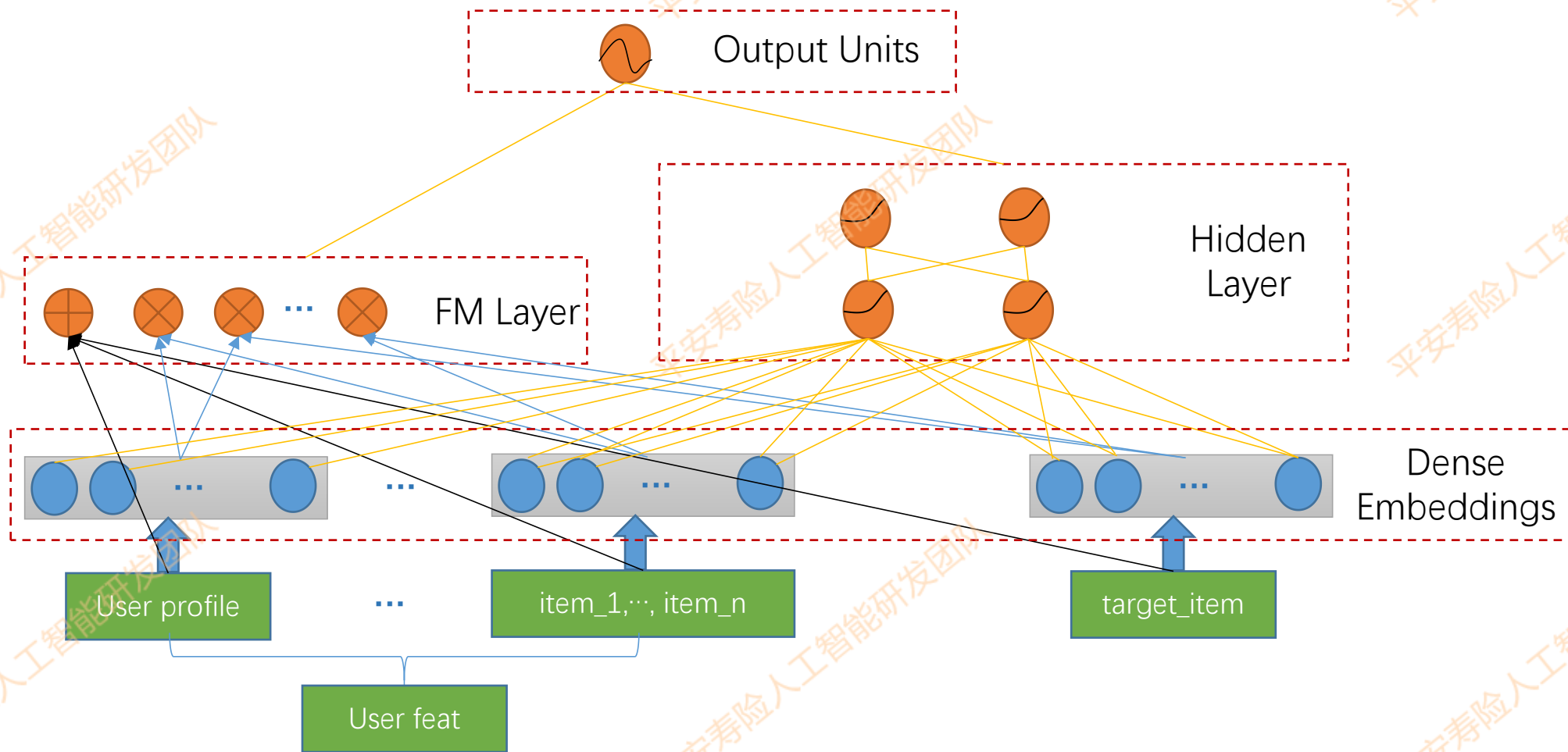
- 基于特征embedding的DNN模型
 - FNN、PNN
 - 只关注了高阶交叉特征，忽略一阶特征

- DeepFM模型

$$y = \sigma(y_{fm} + y_{dnn})$$

- 在FM模型基础上增加dnn部分
 - 通常FM只有二阶特征交叉
 - dnn部分表达更高阶的交叉特征

DeepFM模型



DeepFM模型效果

育儿圈DeepFM模型

点击最大长度	embedding维度	deep部分隐含单元数	dropout	AUC	Batch_size
15	10	64, 32	0.3	0.7823	512
15	10	64, 64	0.5	0.7880	512

保险圈DeepFM模型

点击最大长度	embedding维度	deep部分隐含单元数	dropout	AUC	Batch_size
15	10	64, 32	0.3	0.6372	512
15	10	64, 64	0.5	0.6061	512

DeepFM模型效果分析

- 两个圈子DeepFM模型表现反差原因分析

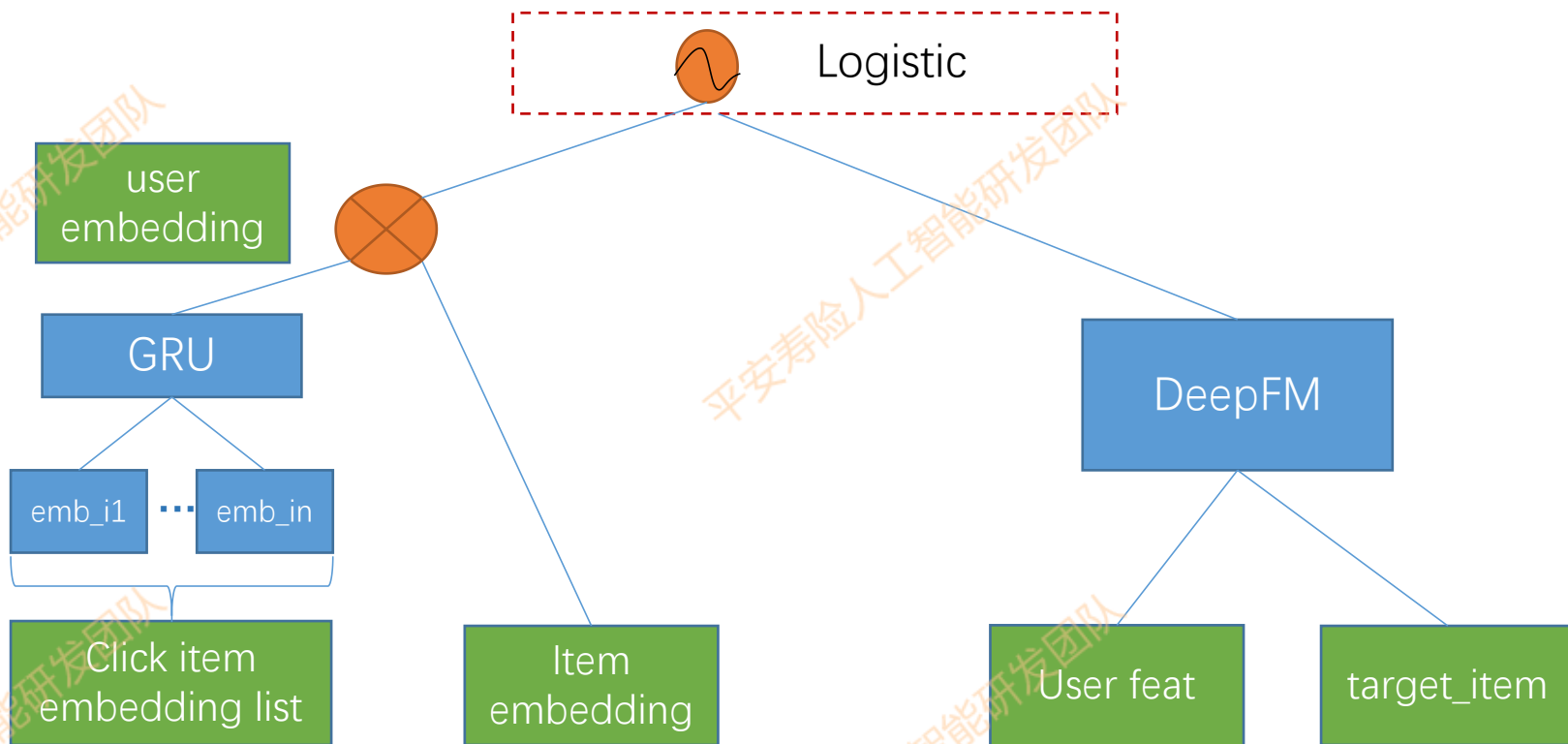
- 保险圈每篇文章平均pv为44.7
- 育儿圈每篇文章平均pv为580
- 需要学习每个特征的embedding
- 文章有较充分的曝光点击，deepfm才能发挥比较好的效果

- 文章语义向量特征不方便纳入

- 稠密的浮点特征
- 单维度与其他交叉表达信息不显著

有效纳入先验信息，提升总体效果？

联合模型DAKUN



联合模型DAKUN

- 结合内容表征与用户行为的模型(DAKUN)
 - Deep Action and Knowledge Union Network
 - gru用户表征模型提取用户隐向量兴趣
 - DeepFM模型提取用户行为特征
 - 内容表征部分能提升冷启动效果

$$y = \sigma(y_{\text{deepfm}} + y_{\text{gru}})$$

联合模型DAKUN效果

保险圈数据

模型	AUC	提升
DeepFM	0.6372	0%
GRU	0.6742	+5.8%
DAKUN	0.6971	+9.4%

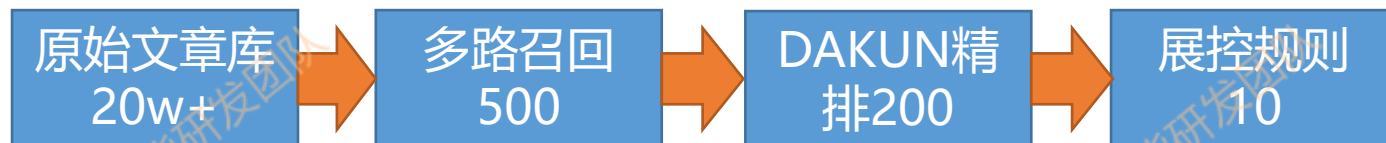
- 联合结构由于单纯任一结构

育儿圈数据

模型	AUC	提升
DeepFM	0.7880	0%
GRU	0.6841	-13.1%
DAKUN	0.7992	+1.4%

总结与展望

圈子feeds



- 内容理解与用户画像是信息流推荐的基础
 - 召回与排序最重要的特征
- 从0到1过程优化召回有更高的收益
- 工程能力是算法提升的基础
 - 召回性能(100->1000)
 - 排序性能, 越复杂模型可排序的条数越低

总结与展望

- 全局召回
 - Faiss向量索引系统
 - 精确表达用户向量，文章向量
 - 阿里TDM
- 多目标排序
 - CTR偏向推荐标题党
 - 点击率、点赞率、收藏率多目标学习(MTL)
- 直接优化set2list
 - 优化目标为最终排序列表
 - 免除人工规则的干预

谢谢大家