

MEIZU

IT大咖说  
不止于技术

# 数据探索

## 魅族技术开放日第七期

● 活动时间:3月18日(周六) 13:30-17:20

活动地点:深圳市南山区科发路1号富利臻大厦3F 思微2.0

联合主办: MEIZU  Flyme msup<sup>®</sup>

MEIZU

IT大咖说  
不止于技术

# 魅族资讯流推荐算法的实践与演进

● 吴梦玲

魅族算法工程师

联合主办: MEIZU Flyme msup<sup>®</sup>



背景介绍



整体推荐框架



实践与演进过程



总结与展望



## 背景介绍

---

# 背景介绍

从目前移动互联网的大环境上看，资讯流是最能给产品带来商业价值的形式之一



V.S.



在线时间足够长、细分度足够高

# 背景介绍

对于Flyme而言：

- 用户阅读资讯内容的最便捷入口
- 整合全系统用户偏好，满足用户对内容的个性化需求





## 整体推荐框架

---

# 整体推荐框架

就因为我前几天搜索了一下“宝宝出牙顺序”，这几天打开网页全屏都是恶心的各种牙齿烂掉的广告...推广告可以，算法能不能靠谱点！你还不如给我推宝宝牙刷靠谱！这体验能好吗？这ecpm能高吗？😓要走的路还很远啊.....

2017年3月10日 10:10



“做人民群众喜闻乐见的资讯推荐”



✓ 准确性

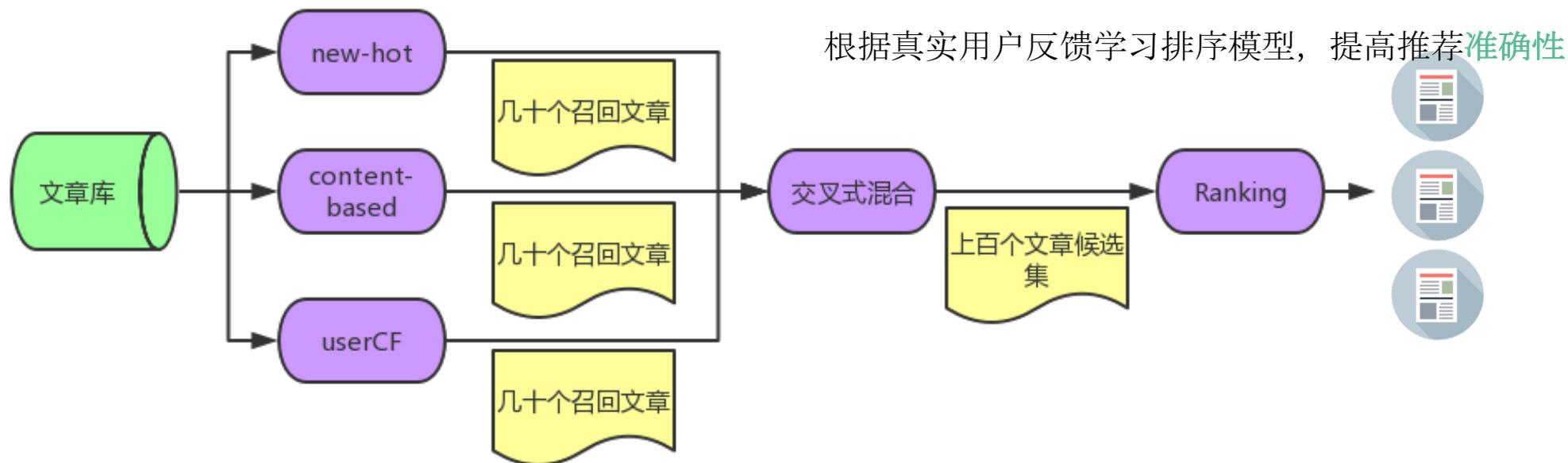


✓ 多样性



✓ 时效性

# 整体推荐框架

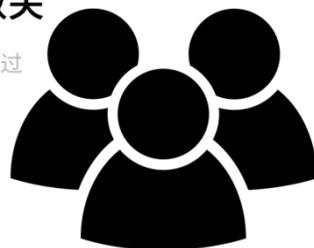


1. 多个召回算法并行组合, 满足推荐结果多样性
2. 召回模型在线更新, 满足时效性

# 整体推荐框架

朴信惠李钟硕代言乐天免税店 中国粉丝称要取关

网易社会 4434人看过



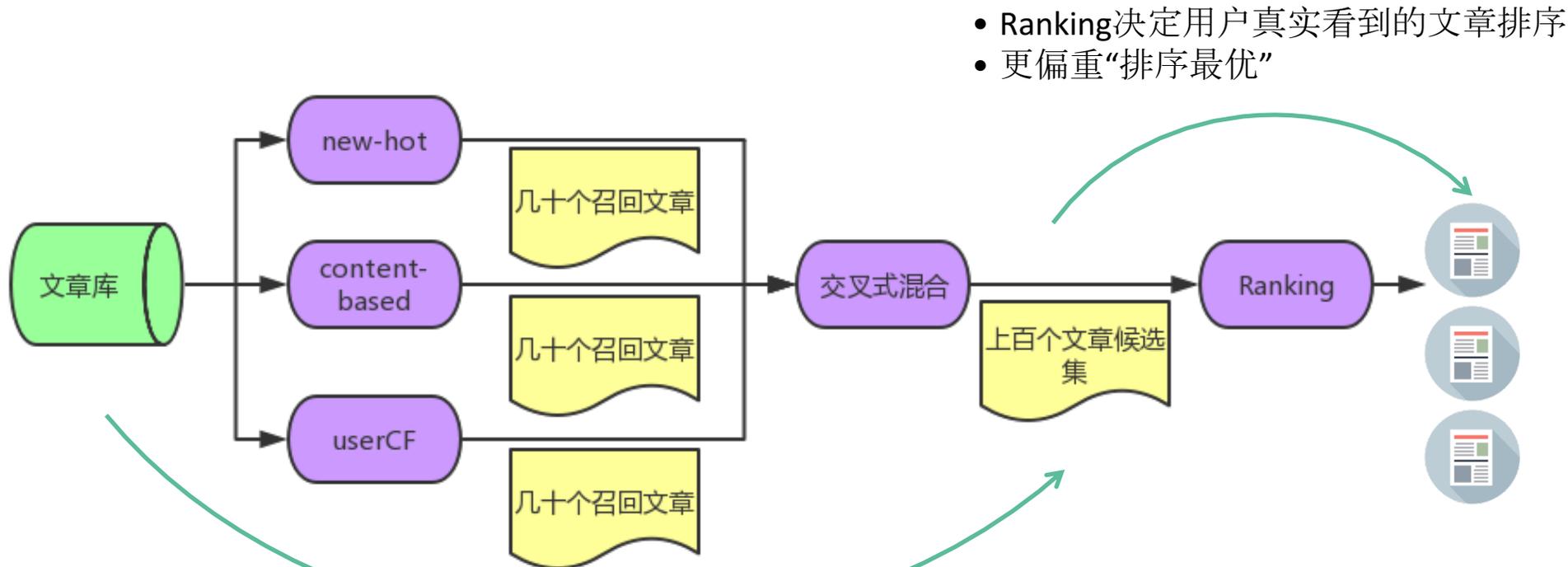
new-hot: 热点新闻+群体偏好

content-based: 个人内容偏好, 特化能力强

userCF: 同类人群的偏好, 相对泛化



# 整体推荐框架

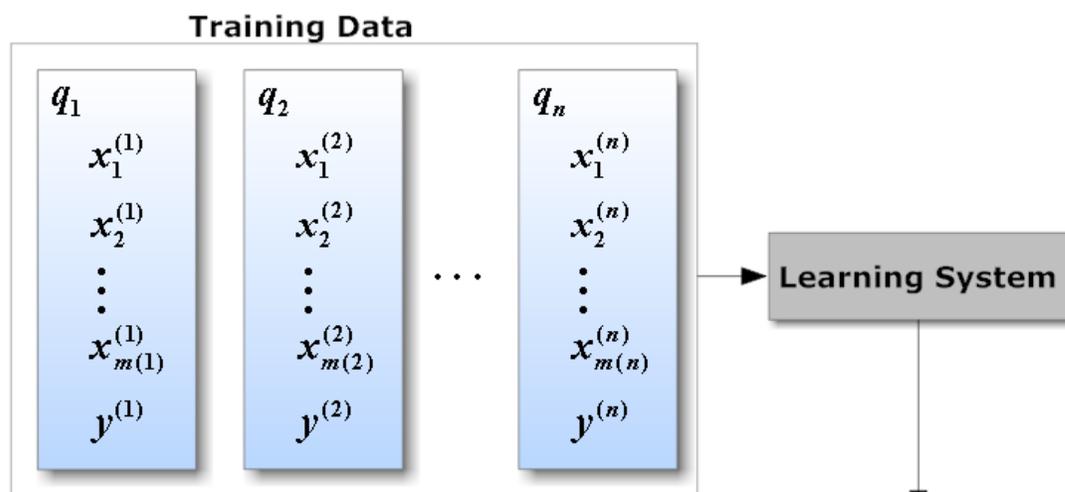


- Ranking决定用户真实看到的文章排序
- 更偏重“排序最优”

- 召回决定用户一次请求可能看到的全部文章候选集
- 更偏重“查全率”

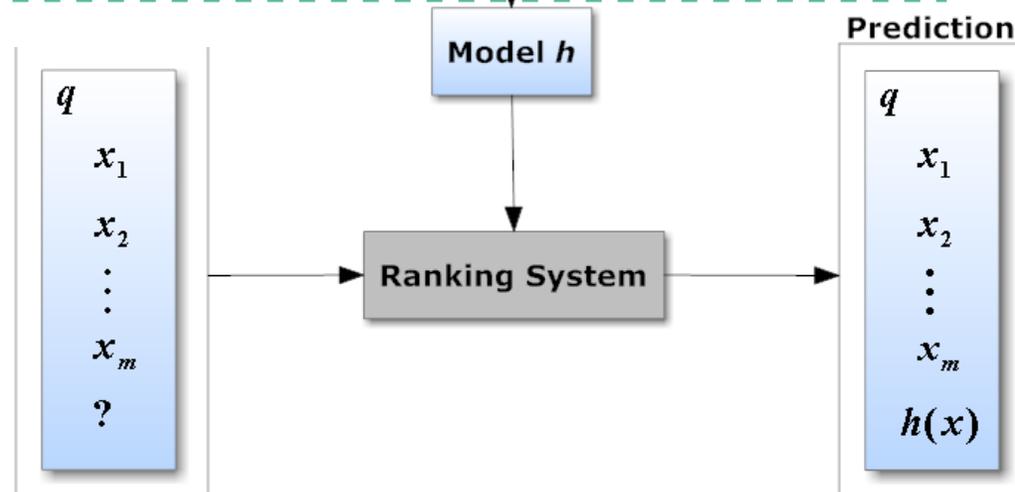
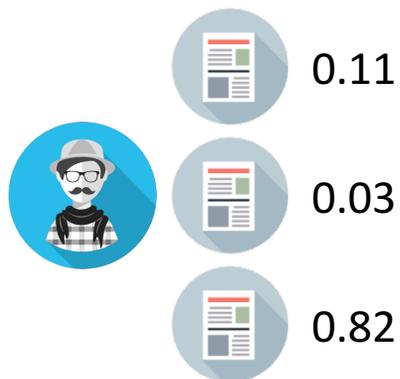
# 整体推荐框架

基于真实反馈数据的Ranking模型：



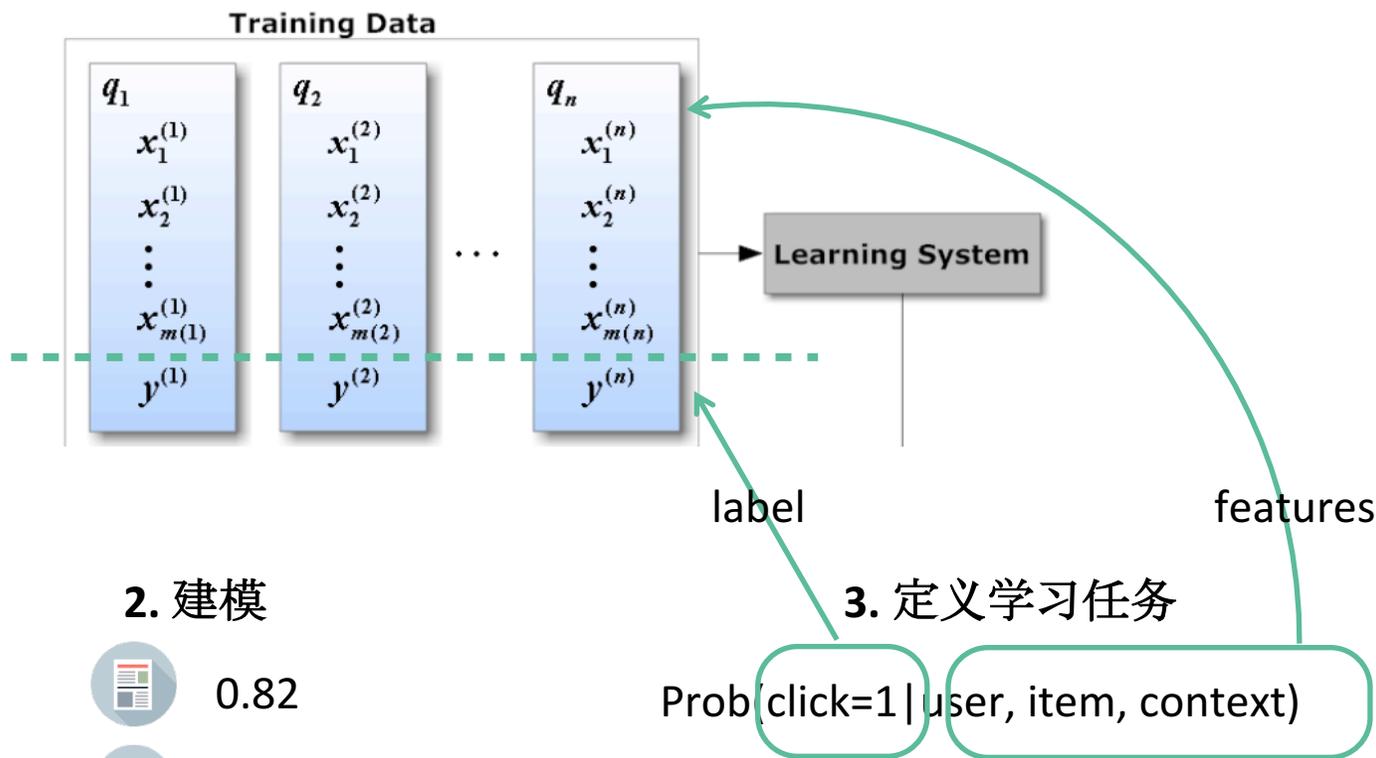
某个晚上11点wifi环境下

在线Ranking:



# 整体推荐框架

Ranking模型离线训练:



## 1. 可量化的业务目标



整体转化率

## 2. 建模

-  0.82
-  0.11
-  0.03

根据文章的预估转化率决定展示顺序

# 整体推荐框架

## 1. label构造

0 河北无端挑衅执勤交警男子被抓获(组图)

国际在线 2176人看过



0 朴槿惠下台退居私邸 支持者上门送玫瑰花

凤凰新闻 9804人看过



1 三星太子李在镕身在看守所，但仍“遥控”三星

IT之家 5640人看过



0 网民消费维权报告出炉 北上广深市民维权意识领先

凤凰新闻 3986人看过

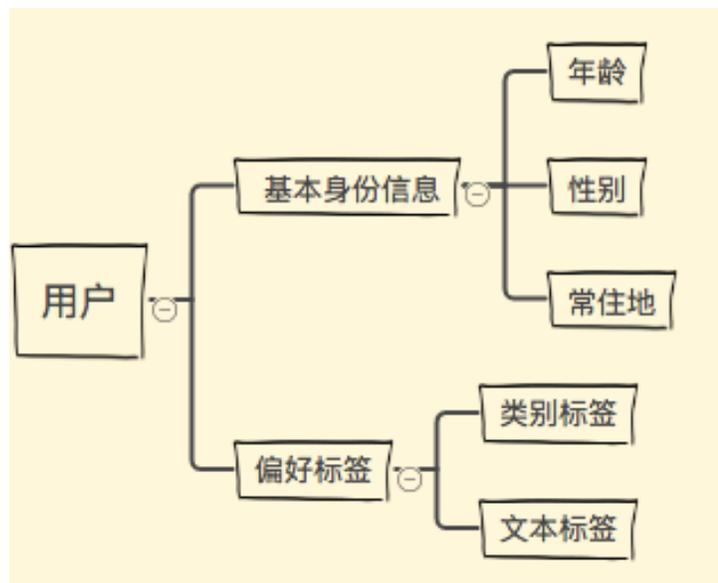
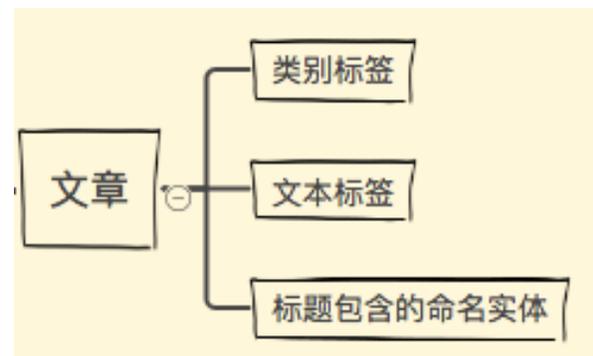
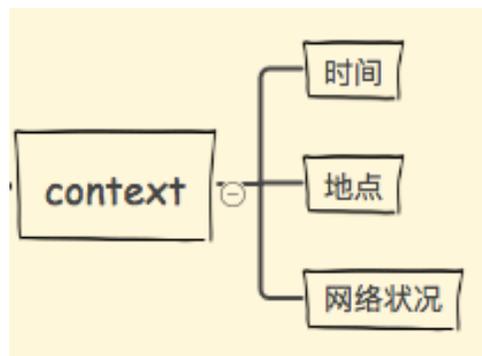


0 在这些实例中，学会用动效赋予你APP生命力

优设 7819人看过



## 2. features构造



LR

# 整体推荐框架

## 18.6%

李克强：完全有能力扩大就业 不允许出现大规模群体性失业

凤凰新闻 1.1万人看过



外媒称中国多地房价暴涨戛然而止 涨幅得到控制

网易财经 1万人看过



她感染H7N9捡回一条命，只因做对这件事

头条热点 1.8万人看过



广州知名火锅店惊人内幕 一锅底料重复利用560次



谷歌Play Store上线五周年 热销内容中国用户也不陌生

腾讯科技 2151人看过



亲子活动大蛋糕怎么切？童玩儿要深耕电商环节

芥末堆 2682人看过



济南有个“90后”供暖管家，工作俩月就当站长！

凤凰新闻 7234人看过



谈谈数学检验分析在产品上线测试中的应用



03

## 实践与演进过程

- 召回模型在线更新
- Ranking模型演化



## 召回模型在线更新

---

# 召回模型在线更新

新文章冷启动



user_1	item_11, item_12, ..., item_1m
user_2	item_21, item_22, ..., item_2m
user_3	item_31, item_32, ..., item_3m
...	...
user_n	item_n1, item_n2, ..., item_nm

# 召回模型在线更新

## 启动页



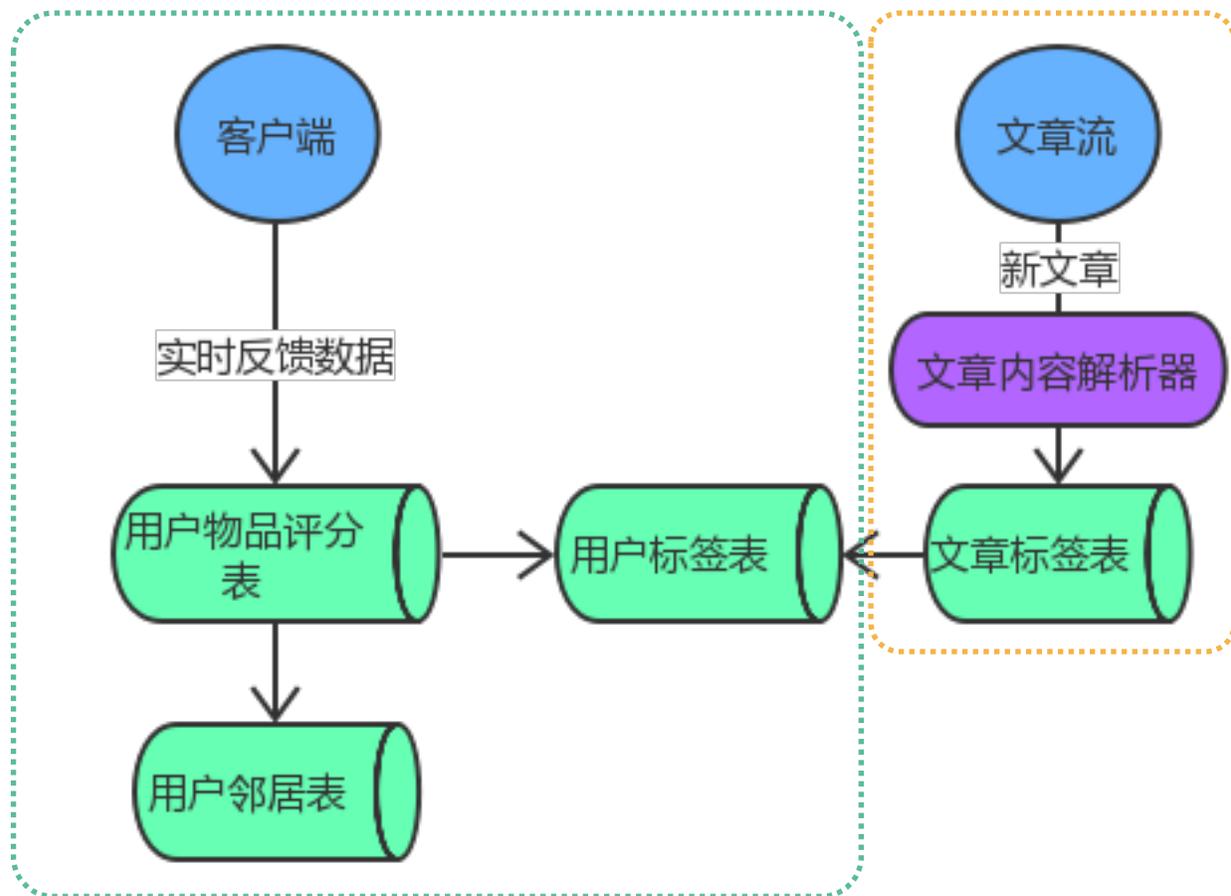
随时间产生新行为，反映偏好



新用户

user_1	item_1, item_2, ..., item_m
user_2	item_1, item_2, ..., item_m
user_3	item_1, item_2, ..., item_m
...	...
user_n	item_1, item_2, ..., item_m

# 召回模型在线更新



文章profile在线更新  
解决新文章冷启动

用户profile在线更新  
捕捉新用户冷启动后的实时偏好

# 召回模型在线更新

3月12号22点某个新用户看到的第一个启动页：

苏醒放飞自我庆巴萨逆转，要给球迷发一万红包！鹿晗李宇春吴亦凡都是铁杆粉丝，周杰伦竟是个“假”球迷？

中国足坛昔日五大遗憾未能留洋的希望之星

曹操至少有两次统一天下：为何没能抓住机会

男医生弯腰低头与病患家属交流被拍 网友：很温暖

接下来10分钟浏览的文章：

赵薇农村买西瓜被强摘眼镜确认身份，立马翻脸讨瓜！

麻麻我是捡的？

中国有微信 支付宝 国外有哪些移动支付方式

之后通过算法召回的文章：

赵薇生日第一个送祝福的竟是张萌 随后张瑶等也深夜送上祝福

四大女神组资本牌局：杨幂运气好，赵薇路子野，范爷、诗爷诈胡

女人干哪些行业注定单身？没想到这一行上榜了

如何用微信快速给信用卡还款？简单几步轻松搞定！

iPhone 8用不用曲面屏还两说 苹果可能要换想法了

林心如晒照庆赵薇41岁生日：宝贝儿，永远爱你

鲜肉网红直播被呛 怒吃洗面乳网友吓歪

李明启“老戏骨”精湛演技“百变老太”生活照看上去好慈爱老太太



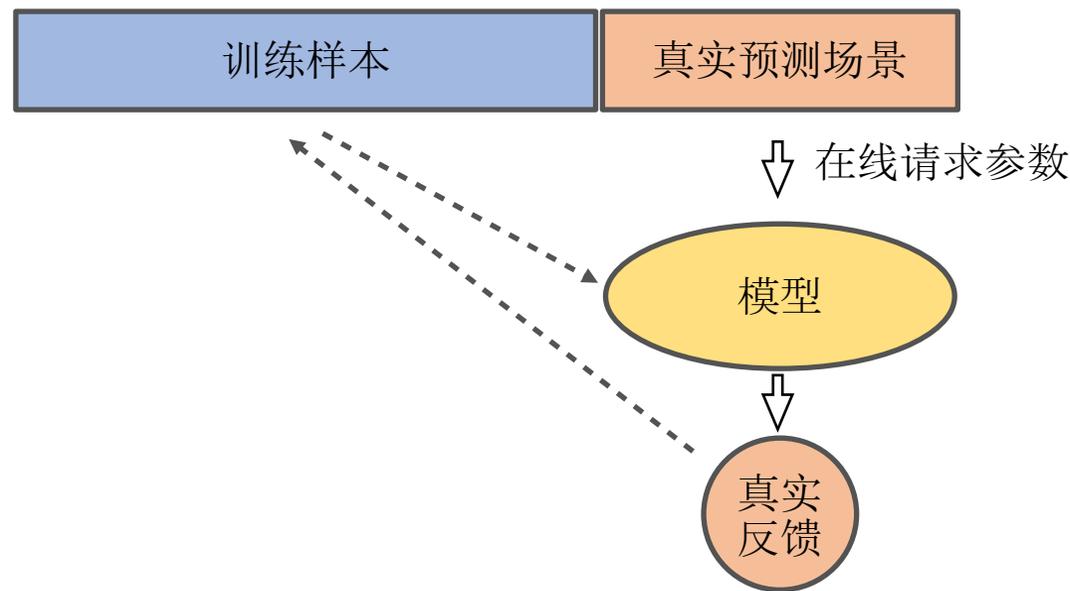
## Ranking模型演化

---

# Ranking模型演化

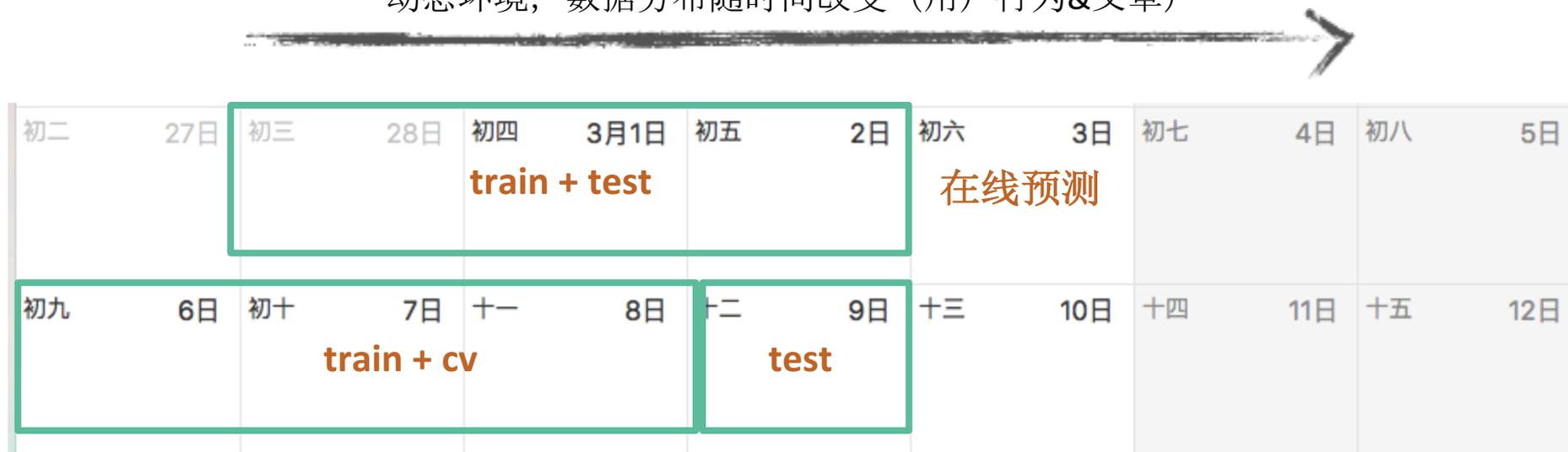


- 训练数据的label和特征与线上环境对齐
- 使得离线实验效果尽量与在线效果正相关



# Ranking模型演化

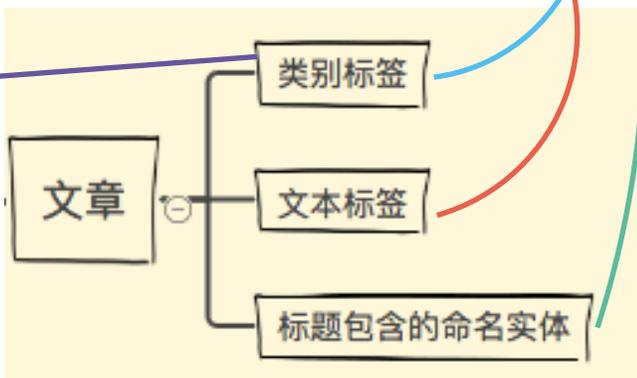
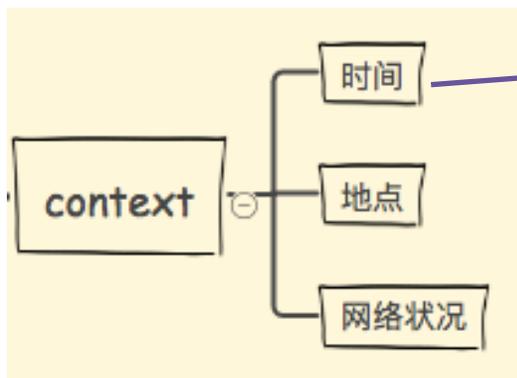
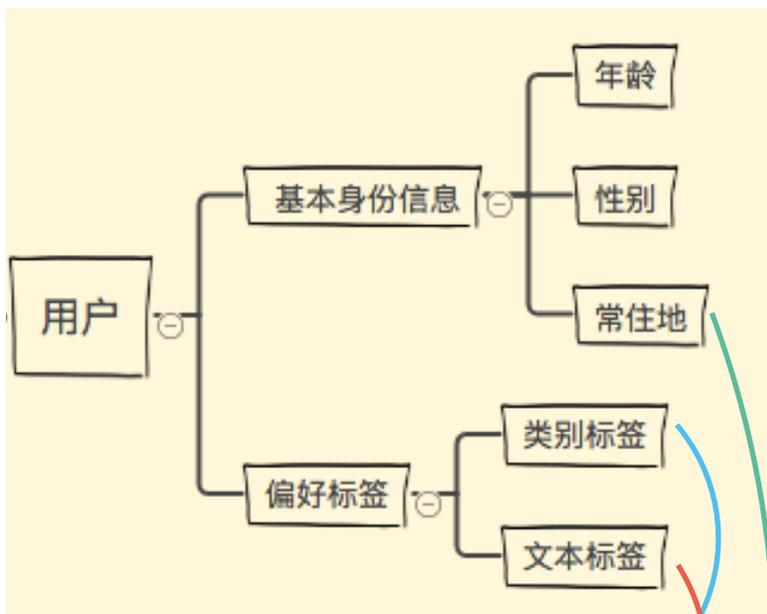
动态环境，数据分布随时间改变（用户行为&文章）



1. 用连贯的之后若干天作为test集，检验模型泛化能力
2. 拉大训练样本时间周期，减轻过拟合

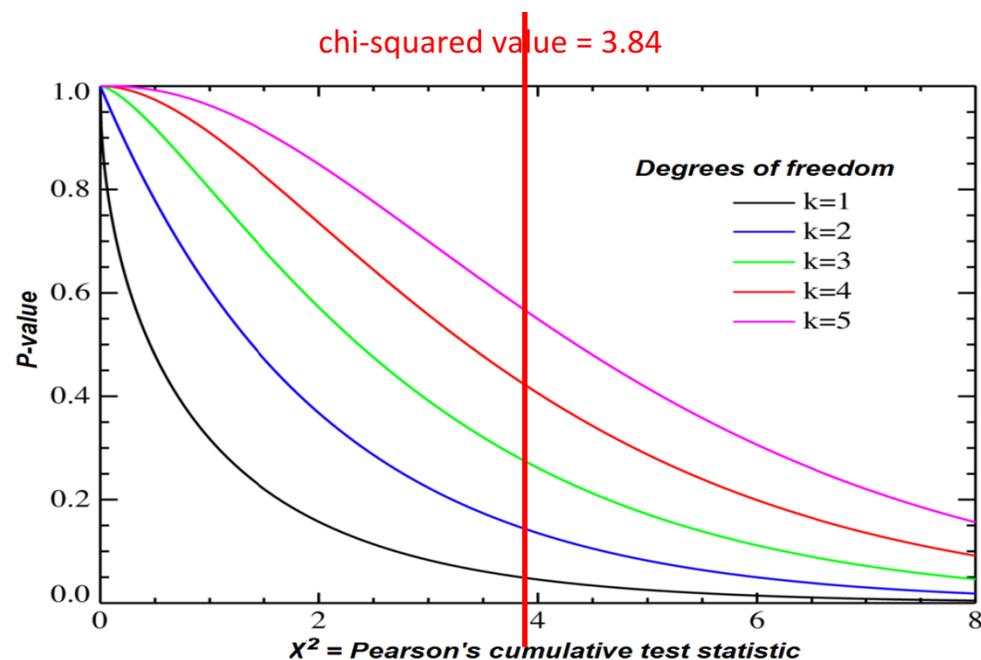
# Ranking模型演化

## 1. 二阶特征组合



## 2. 特征抽取

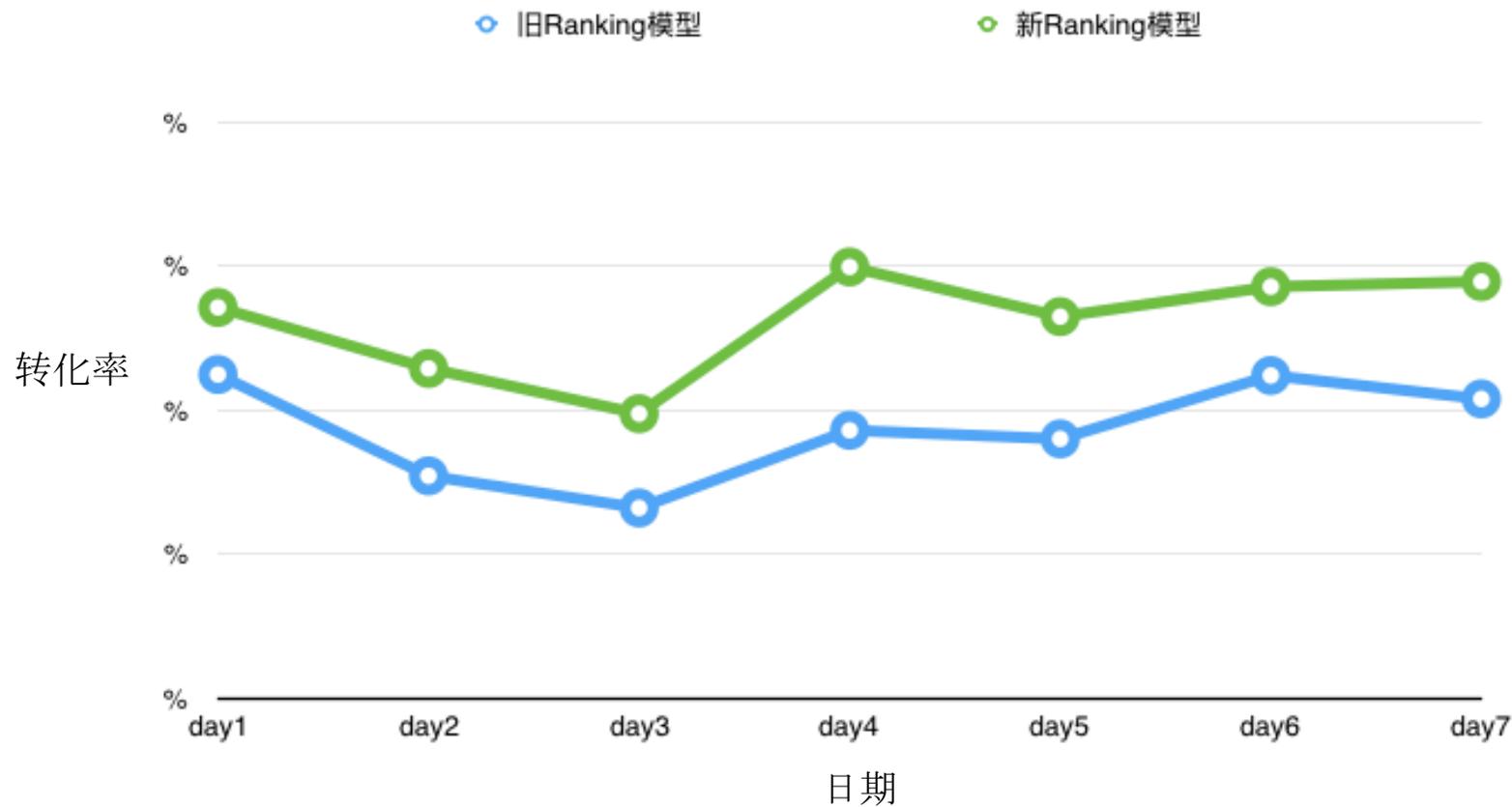
- 只在正样本/负样本中出现
- 相关性低



(图片来自维基百科)

# Ranking模型演化

12.86%



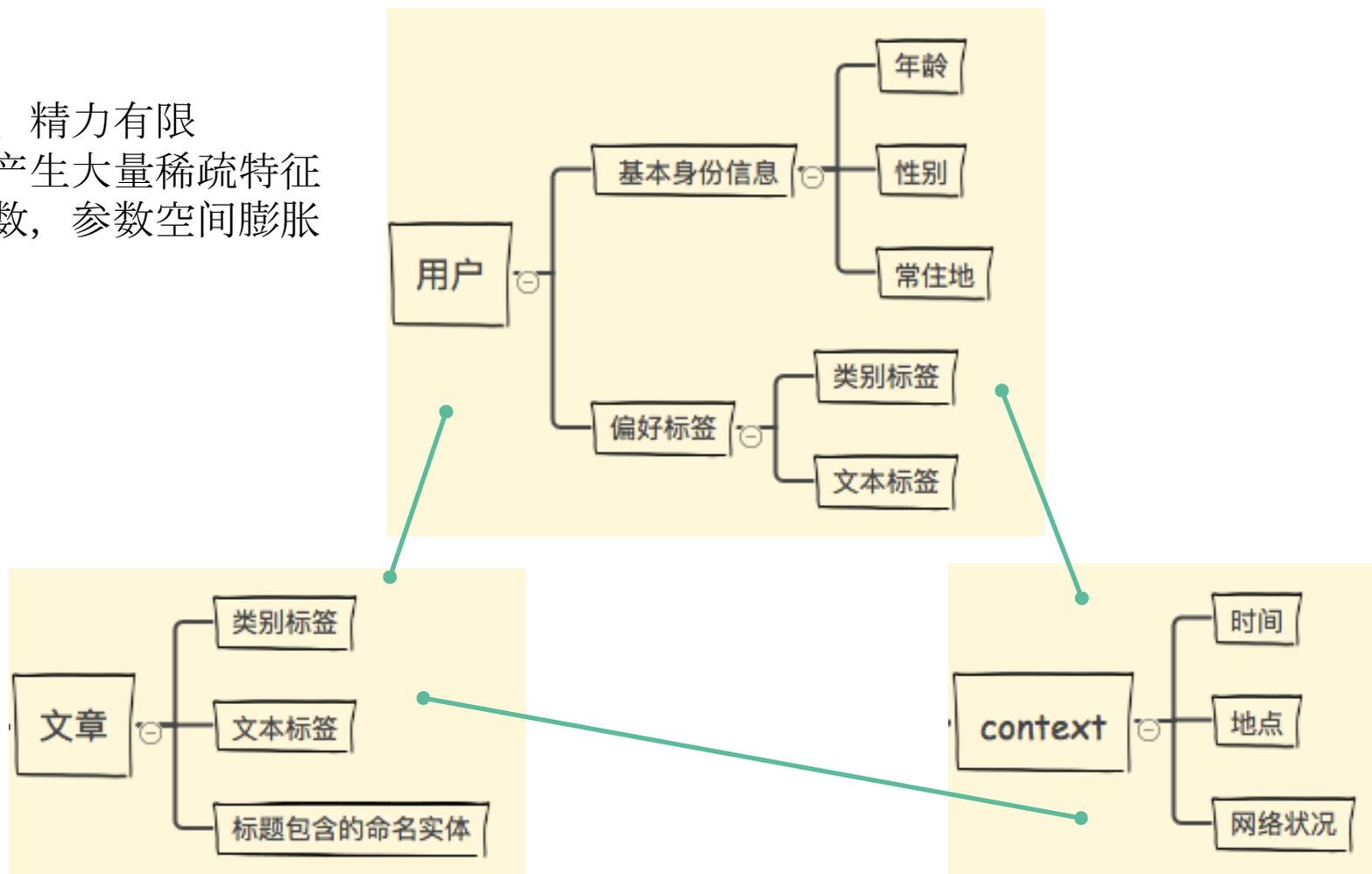


## 总结与展望

---

# 总结与展望

1. 在特征构造上人的先验经验、精力有限
2. 离散特征的高阶组合天然会产生大量稀疏特征
3. 每个高阶项独有一个权重参数，参数空间膨胀

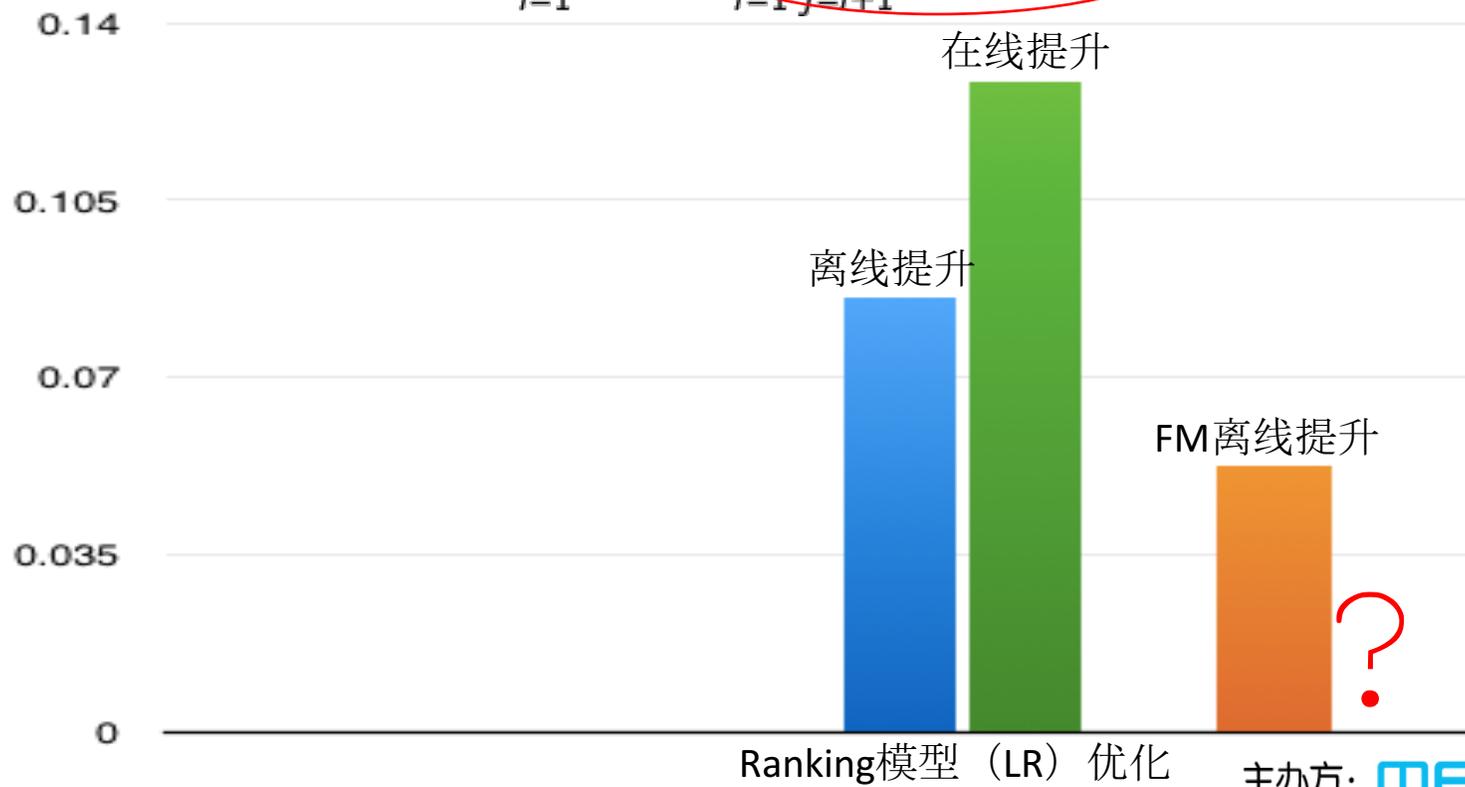


# 总结与展望

$$y = \text{logit}(w_0 + \sum_{i=1}^p w_i x_i + \sum_{i=1}^p \sum_{j=i+1}^p w_{ij} x_i x_j)$$

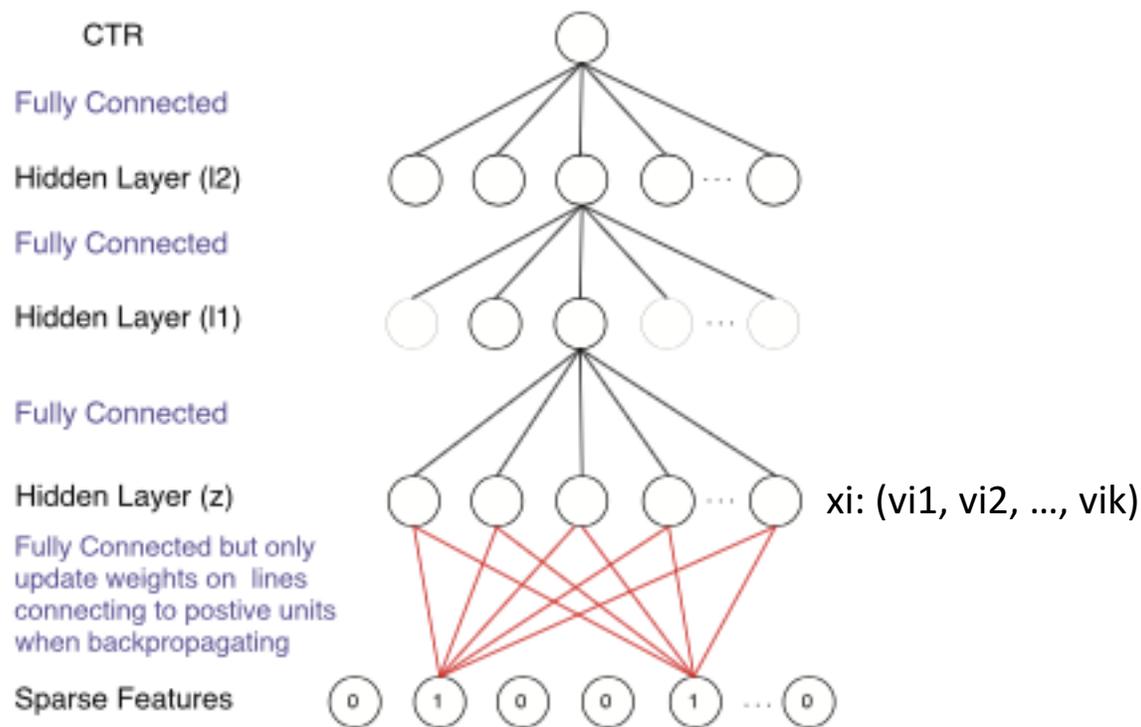
↓ xi: (vi1, vi2, ..., vik)

$$y = \text{logit}(w_0 + \sum_{i=1}^p w_i x_i + \sum_{i=1}^p \sum_{j=i+1}^p \langle v_i, v_j \rangle x_i x_j)$$



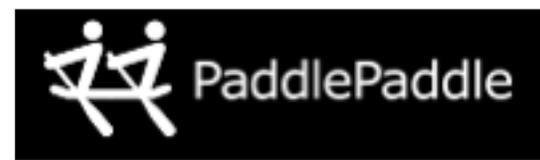
# 总结与展望

将FM得到的连续性因子向量作为DNN的其中一层



(a) SNN Architecture

(图片来自Weinan Zhang的论文Deep Learning over Multi-field Categorical Data)



# Thank you!

联合主办: MEIZU  Flyme msup<sup>®</sup>