

# 智能推荐演变之路

---

阿里巴巴-创新事业群-夸克智能搜索

王 跃

2019.8.17

01

## 概览

02

## 召回

用户行为分析  
标签召回  
知识图谱召回  
向量召回

03

## 排序

基础相关性模型  
点击率预估模型  
MAB

www.aicug.cn  
AICUG 人工智能技术社区

# 概览-场景介绍

## ➤ 目标:

- 导流, 流量=钱
- 内容循环消费

## ➤ 输入:

- Sug
- 预置搜索词
- 搜索发现

## ➤ 搜索结果页:

- 文字推荐
- 为您推荐
- 实体推荐、app推荐、小说推荐
- 相关搜索

## ➤ 内容页:

- H5底部推荐
- 转码页推荐



# 概览-技术大图



01

概览

02

召回

用户行为分析  
标签召回  
知识图谱召回  
向量召回

03

排序

基础相关性模型  
点击率预估模型  
MAB

www.aicug.cn  
AICUG 人工智能技术社区

# 召回-用户行为召回

## ➤ 多样性优化

- 针对不同类型的行为进行权重区分
- 对session中最后一次搜索做加权

## ➤ 时效性优化

- 使用X天挖掘日志,中高频挖掘相似query使用动态窗口
- 相似度计算按时间衰减

## ➤ 稀疏优化

- 基于文本的泛化: title相似的doc共享推荐list
- 基于行为的泛化: 相似query共享推荐list

## ➤ 收益

- UrlNextQuery个数 $\geq 5$ 的比例由x%提升到x%
- 搜了又搜在query推荐中覆盖率由x%提升到x%
- 相关搜索Ctr提升x%,通用文字推荐Ctr提升x%

数据对外服务 (Athena)

多样性优化

时效性优化

稀疏优化

搜了又搜  
(NextQ)

搜了又看  
(Url PreQ)

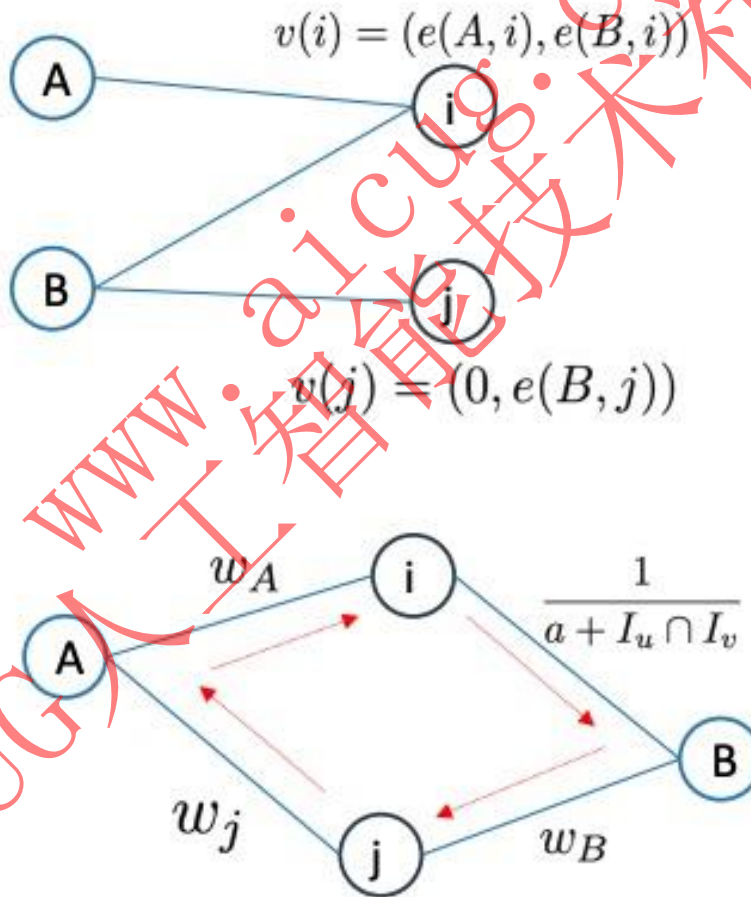
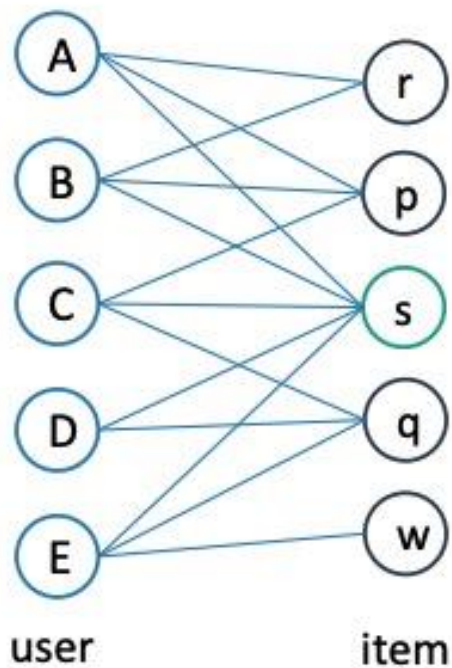
看了又搜  
(Url NextQ)

相似  
Query&Url

用户session日志

铜矿日志(用户搜索,浏览日志)

# 召回-用户行为分析



$$r(i, j) = \frac{v(i) * v(j)}{|v(i)| * v(j)}$$

collaborative filter

$$s(i, j) = \sum_{u \in U_i \cap U_j} \sum_{v \in U_i \cap U_j} w_u \cdot w_v \cdot \frac{w_j}{\alpha + |I_u \cap I_v|}$$

where  $w_u = \frac{1}{(\alpha_1 + |I_u|)^{\beta_1}}, w_j = \frac{1}{(\alpha_2 + |U_j|)^{\beta_2}}$

Swing

# 召回-标签召回

## ➤ 目标

- 构建标签库
- 帮助用户准确及时获取自己关注信息的动态

## ➤ 问题&难点

- 随意性强，口语化
- 标签种类多样，资源库标签覆盖不足

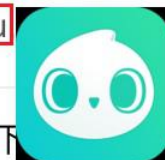
## ➤ 解决方案

- 标签挖掘
  - 人工运营
  - Pattern挖掘
  - 远程监督模型挖掘(BI-LSTM CRF模型)

什么软件拍照带耳朵的\_百度知道

1个回答 回答时间: 2017年02月05日

现在特别火的就是faceu 萌猫耳朵



带耳朵卖萌的拍照软件下

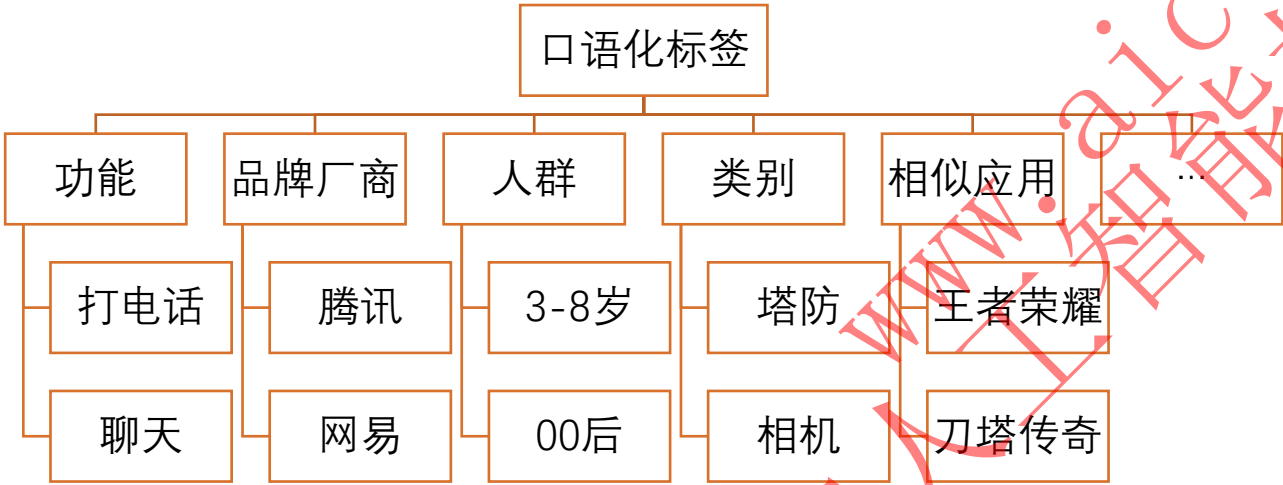
可以戴着猫耳朵什么的动的拍照软件是啥

最近很流行的一个照相软件是什么软件 有很...



# 召回-标签召回

➤ 标签体系



# 召回-知识图谱召回

## ➤ 算法PK

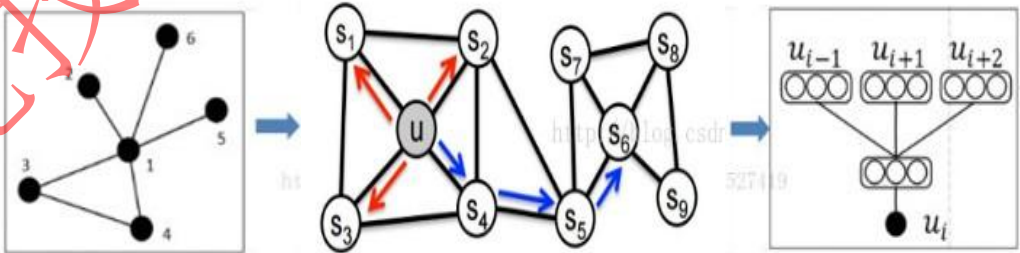
	文本建模	知识表示	网络关系
典型代表	LDA, Doc2vec	tranE, transH, transD, transR	DeepWalk, Node2Vec, SNDE
使用特征	文本	关系	关系
关系敏感度	低	高	中
网络稠密要求	无	高	中

## ➤ Node2vec

- 随机游走
- 深度优先&广度优先
- Skip-gram

## ➤ 优化

- 数据增广：用户行为数据、百科超链接
- 文本信息embedding



# 召回-向量召回

## ➤ 目标

- 捕捉Query和推荐结果之间的语义相关性

## ➤ 样本选择

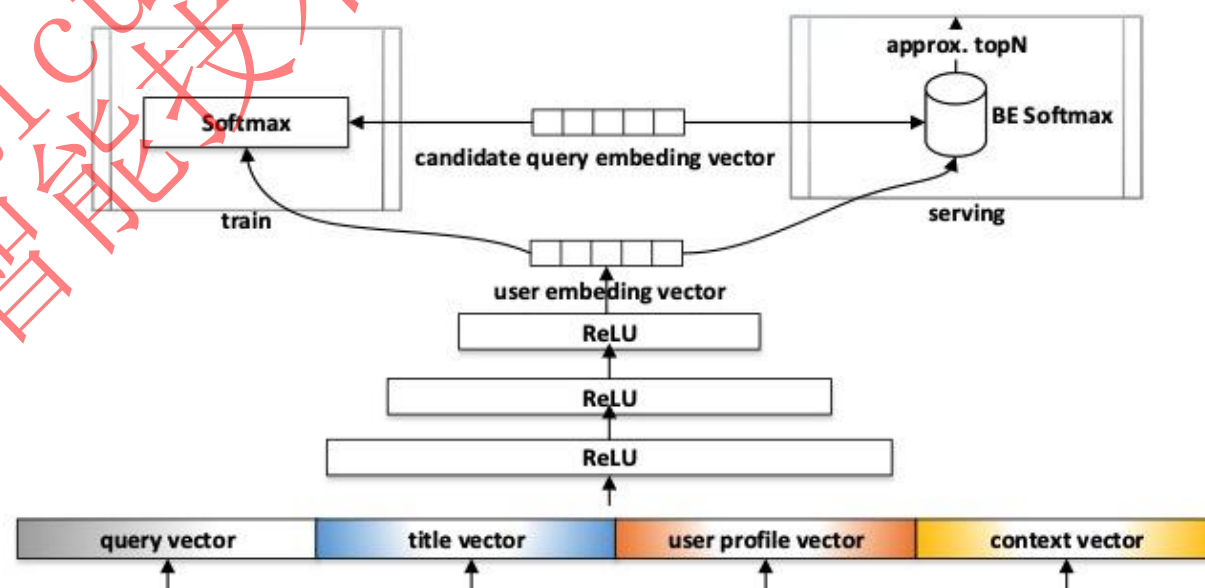
- 同Query下有点击作为正样本，无点击作为负样本
- 归一化操作：超低频删除，超高频subsampling
- 样本比例，正负样本1: x

## ➤ 模型特征

- Query文本特征：检索切词、语义切词
- 用户画像特征：年龄、职业、性别
- 实时信息特征：时间、网络、城市

## ➤ 模型参数

- 文本特征提取：Bi-Lstm, cnn, dnn, bow
- 激活函数：relu, tanh
- Sample loss: nce\_loss, sampled\_softmax\_loss



# 召回-向量召回

输入数据	召回数据	真实搜索Query
[add] act=6:search snid=11:19981884119 utdid=24:WiS8CF7zuLcDAKC/Ptbfeu3M net=4:wifi brand=5:Meizu queries=36:眼睛上眼皮里面长了个疙瘩 last_query=36:眼睛上眼皮里面长了个疙瘩 last_query_ts=4:25.0 titles=0: last_title=69:上眼皮里面长个疙瘩好多天了眼睛很不舒服怎么回事 last_title_ts=69:上眼皮里面长个疙瘩好多天了眼睛很不舒服怎么回事 model=7:m2 note fr=7:android tm=19:2018-02-23 00:24:39 hour=1:0 city=9:西安市 lat=7:34.2803 lon=8:108.9976 age=2:19 sex=1:M entities=36:4d2cc6b2-7c16-11e7-8238-6c92bf477917 preference=13:小说`电影	麦粒肿 14.4656 眼睛麦粒肿初期图片 14.1656 麦粒肿图片 14.1328 眼睑炎 14.0408 散粒肿 13.8903 麦粒肿最初期的图片 13.7273 针眼 13.4677	麦粒肿图片

- 效果分析
- 召回 top1精度x%  
(基线为x%)

01

概览

02

召回

用户行为分析  
标签召回  
知识图谱召回  
向量召回

03

排序

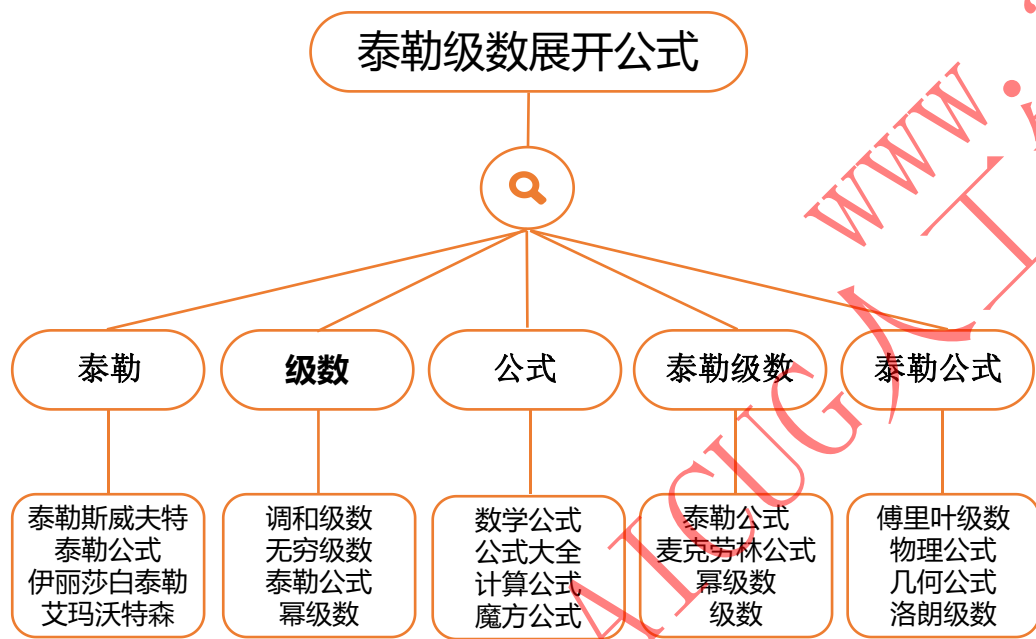
基础相关性模型  
点击率预估模型  
MAB

www.aicug.cn  
AICUG人工智能技术社区

# 排序-基础相关性

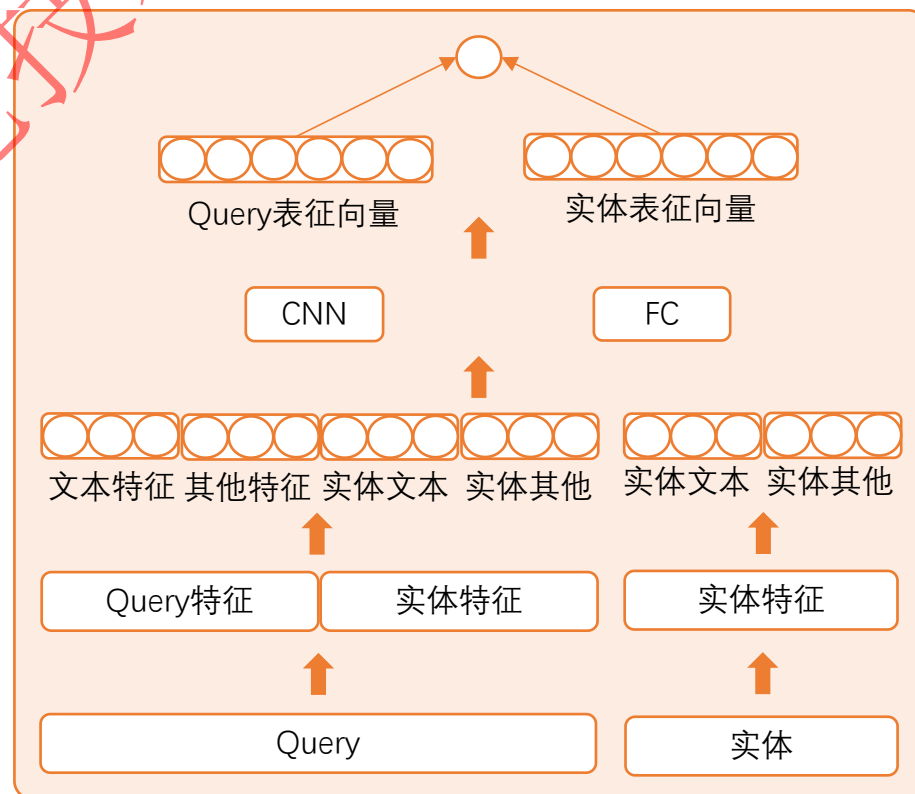
## ➤ 文本相关性

- titleMatch
- docMatch
- phraseMatch
- 同义词
- 语境同义词



## ➤ DeepMatch

- 特征：文本特征、切词特征、实体特征、分类特征
- 模型：分层的DSSM
- 数据：有点击数据正样本，无点击数据负样本



# 排序-基础相关性

Q=刘一男曝光何凯文豪宅	score
2017何凯文直播	0.993785918
新东方2016高考视频百度网盘	0.993198097
何凯文直播	0.992349207
刘一男曝光何凯文豪宅	0.99219048
新东方英语免费视频	0.991401255
网红考研老师视频	0.991224408
何凯文四川	0.99100405
新东方网课视频	0.990925133
文都刘一男个人资料	0.989709318
牛津国际教育	0.989040852

## 相关推荐



[更多泰勒级数展开公式的推荐](#)

## Q=网球选手李娜

### 相关人物



## Q=歌手李娜

### 相关人物





# 排序-CTR预估

## ➤ 目标

- 点击率最大化

## ➤ 策略

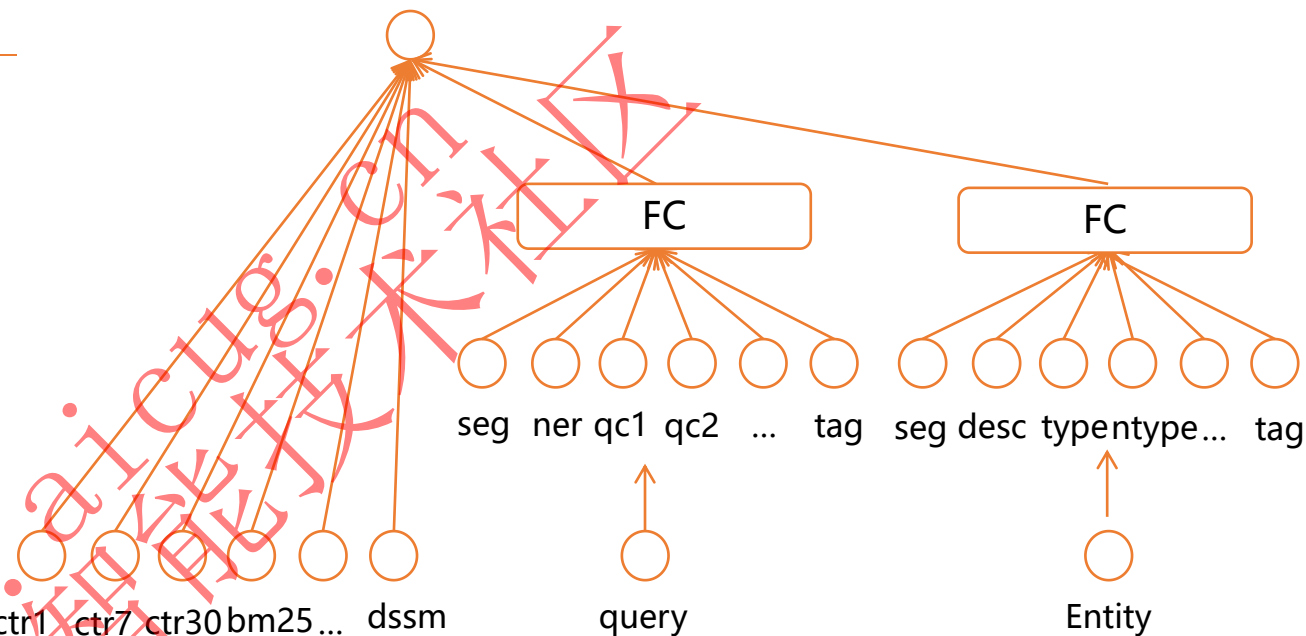
- 成熟算法匹配业务场景
- 匹配业务场景，进行模型设计

## ➤ 模型

- 样本选择：有点击为这样本，无点击为样本，设置pv阈值提升样本置信度
- 样本规模：一期一亿正样本，每天增量叠加
- 模型更新，每天增量训练，提升时效性
- Wide: ctr类特征，文本匹配类特征，召回算法类特征
- Deep: 文本特征，实体特征，分类特征，不同特征各自取embedding维度
- 特征优化：单特征auc实验

## ➤ 收益

- A/B test对比基线ctr提升x%





# 排序-MAB

## ➤ 问题

- 如何平衡探索和利用已有知识

## ➤ 模型选择-MAB

- **UCB**

$$X_i = X_i + \sqrt{\frac{2 \ln n}{n_i}}$$

$X_i$ 是T轮迭代的平均收益,  $n$ 是总执行次数,  $n_i$ 是当前Query执行的次数

- **Thompson Sample**

$$P_{j,t} \sim \text{Beta}(S_{j,t} + 1, F_{j,t} + 1)$$

$S_{j,t}$ : 点击次数,  $F_{j,t}$ : 未点击次数

## ➤ 在线实验:

- TS超参数选择

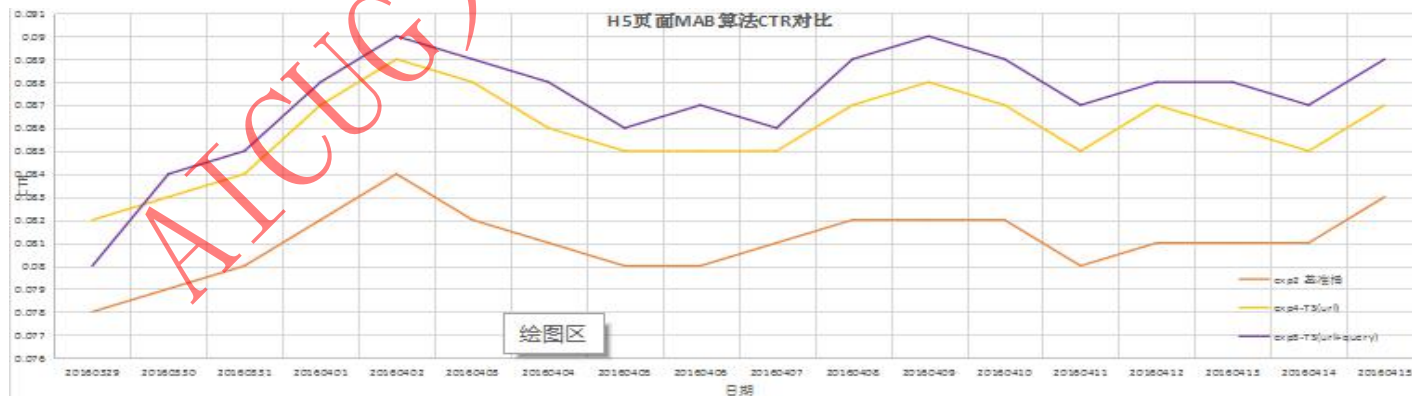
Succ = click + wide&deep score

Fail = 1 + pv - click + 平滑因子 \* (1 - ctr)

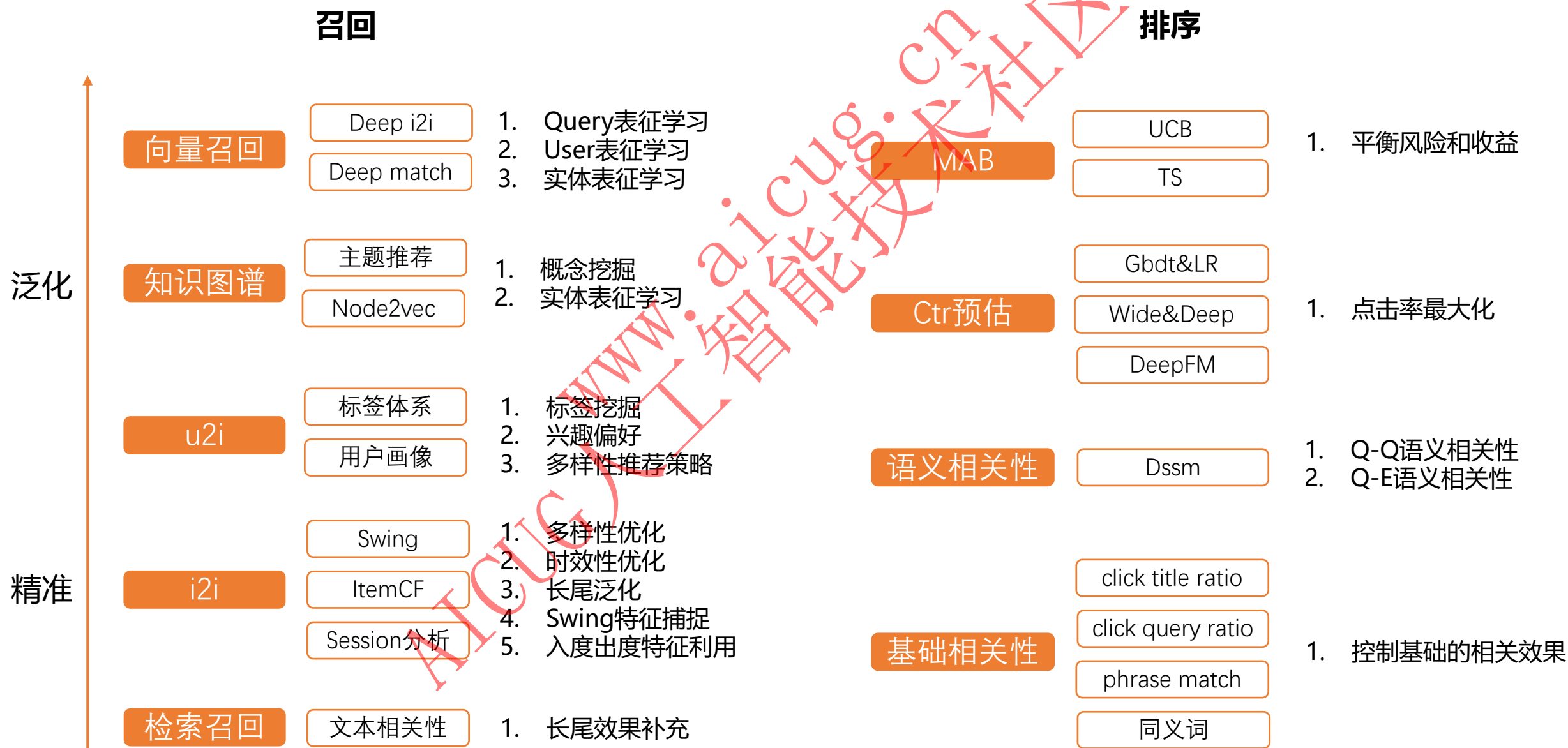
- TS使用Beta随机, 不会困在早期的错误里, 鲁棒性更强
- UCB: 不仅仅关注收益, 还关注Arm的执行次数, 收敛速度相对要慢
- 在线对比UCB收敛速度比TS要差一下

## ➤ 收益:

- url推荐场景下在线ctr提升x%
- 增加实时日志反馈, ctr提升x%



# 小结



# Thanks

---

AICUG 人工智能技术社区  
www.aicug.org.cn