



## 十年架构 成长之路

**SACC** 第十届中国系统架构师大会  
SYSTEM ARCHITECT CONFERENCE CHINA 2018

2018年10月17-10月21日 北京海淀永泰福朋喜来登酒店



# 短视频推荐系统实践

搜狐视频 李修鹏



第十届中国系统架构师大会  
SYSTEM ARCHITECT CONFERENCE CHINA 2018



# 目录

1

推荐系统架构

2

助力推荐系统成长

3

一些问题思考和实践



十年架构 成长之路



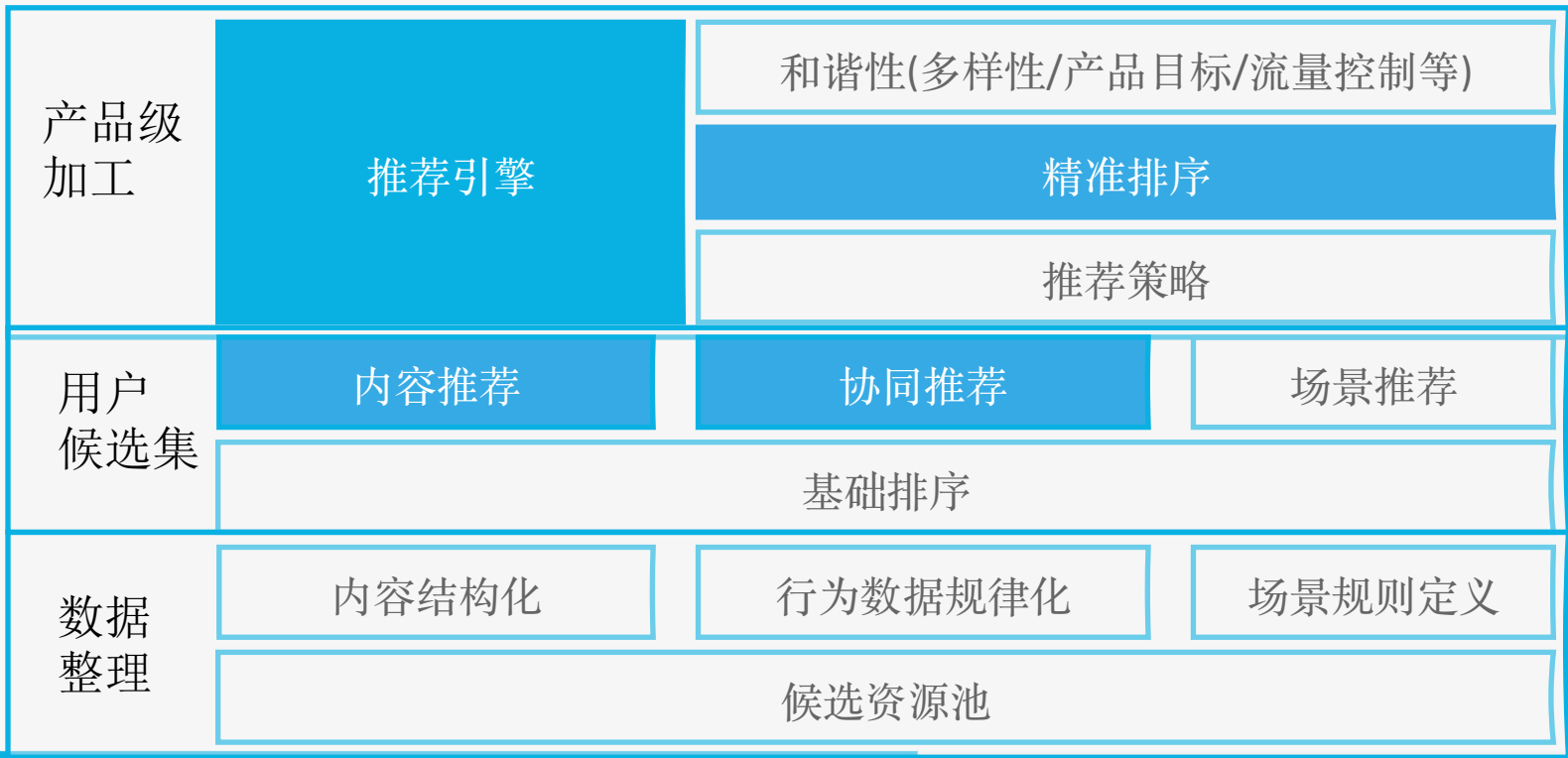
# 推荐系统架构



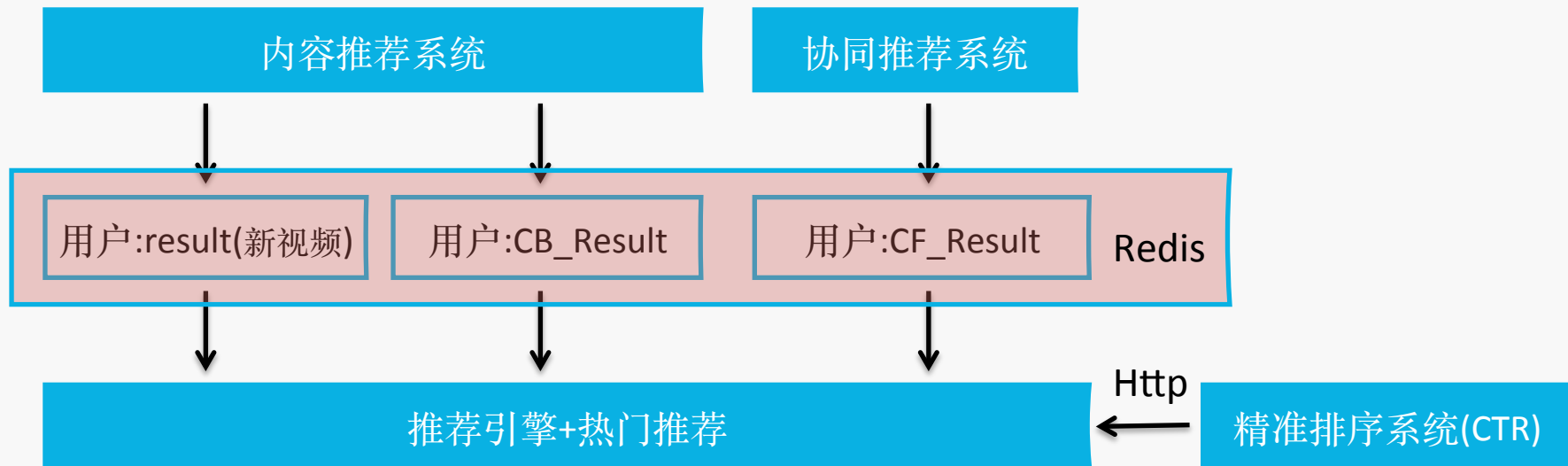
十年架构 成长之路



# ▶ 推荐架构

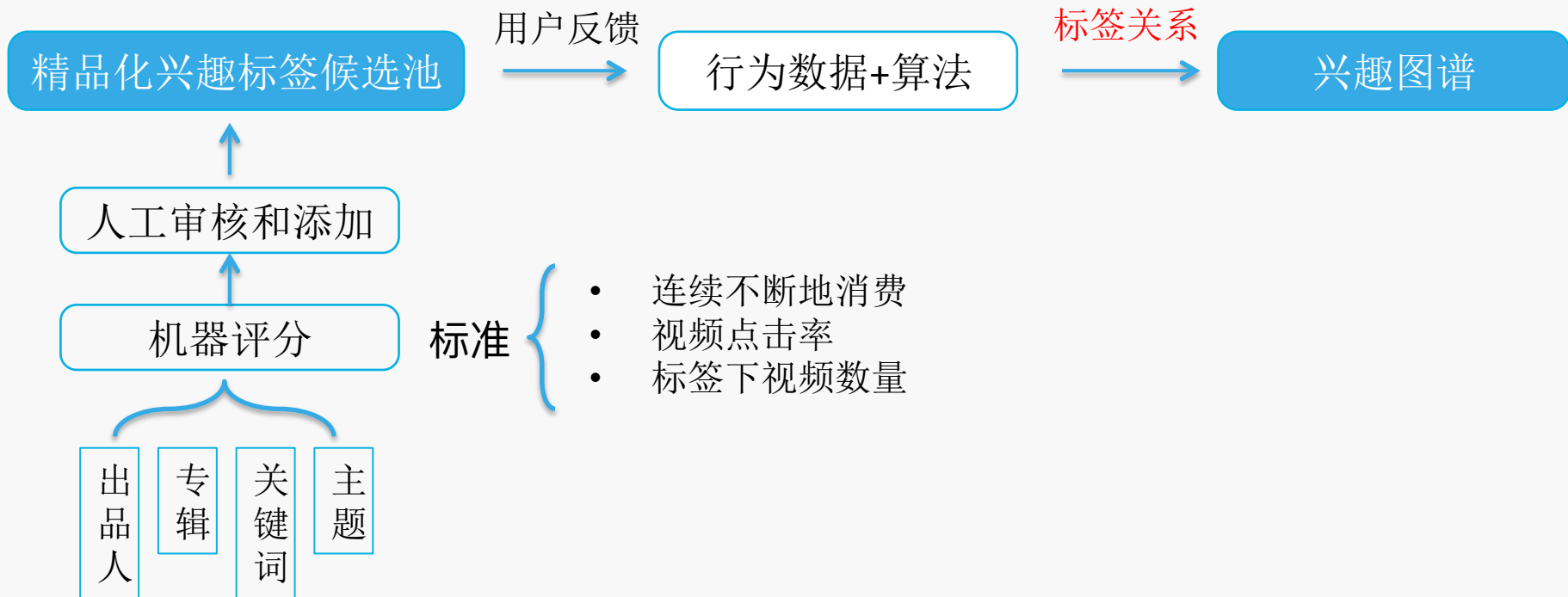


## ▶ 推荐架构

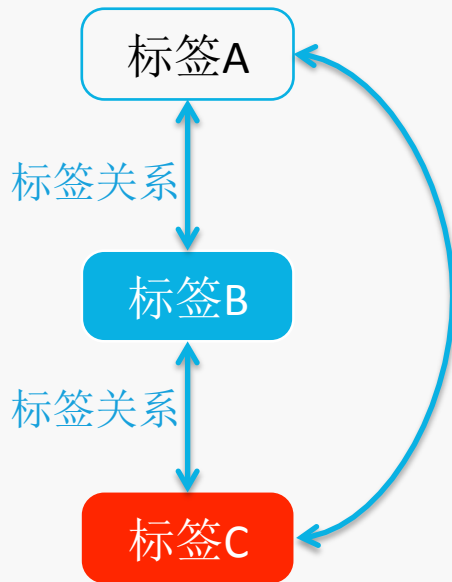
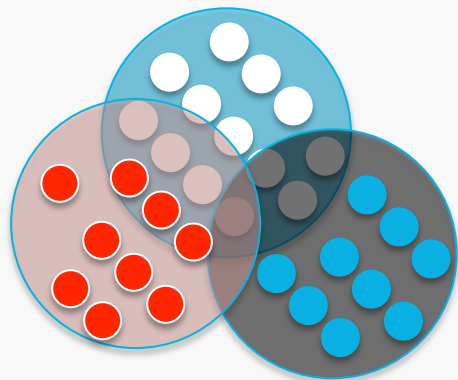


相关度									热度	时效
一级类	二级类	三级类	兴趣图谱	主题模型	关键词	Embedding	出品人	...	E&E	Time
兴趣分类 层级结构 粗			精品化 兴趣标签 (细)	关键词 共现 聚合  图片特征 聚合	人工打 标签  标题提 取	视频标题 embedding <128维> sen2vec			展示/点击  梯度收敛	上传 时间

# 内容推荐系统







标签关系

- 内容相似度:标题、分类等
- 集合的视频观影序列的共现度

# 协同推荐系统

协同基础数据

记忆和归纳得出消费规律

用户行为数据:

	u1	u2	u3	u4	...	uk	...	...	...	um
i1	s11	s12	s13	s14	...	s1k	...	...	...	s1m
i2	s21	s22	s23	s24	...	s2k	...	...	...	s2m
i3	s31	s32	s33	s34	...	s3k	...	...	...	s3m
...	...	...	...	...	...	...	...	...	...	...
ik	sk1	sk2	sk3	sk4	...	skk	...	...	...	skm
...	...	...	...	...	...	...	...	...	...	s1m
in	sn1	sn2	sn3	sn4	...	snk	...	...	...	snm

序列预测

- Item: 视频、分类、标签、主题
- 序列: Item1... Item n观影后 下一个最佳item?

群体效应

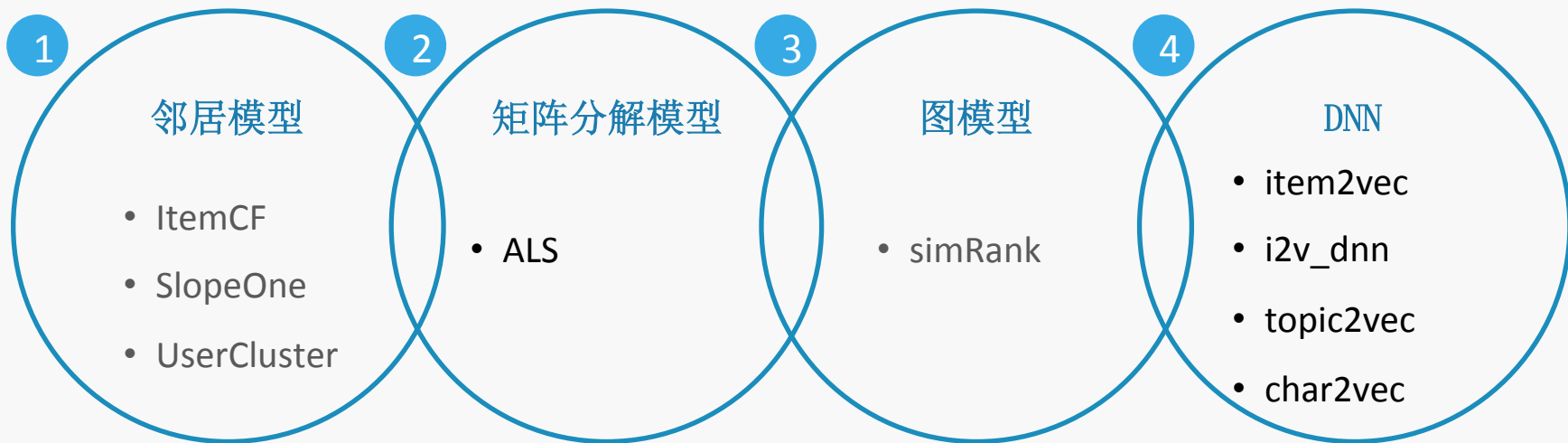
- 兴趣群组(用户)内投票推荐
- 相似兴趣用户组(用户)交叉推荐



SACC

第十届中国系统架构师大会  
SYSTEM ARCHITECT CONFERENCE CHINA 2018





# 协同推荐系统

item2vec

训练数据: 用户观影看成doc  
word: 视频

word2vec

视频embedding

# 协同推荐系统

sen2vec

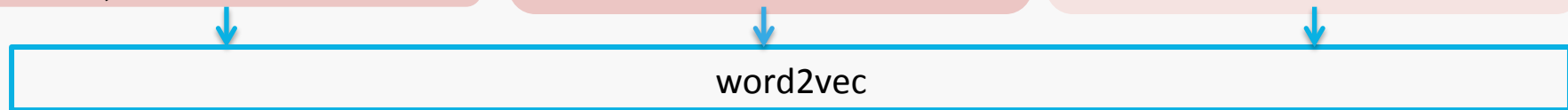
训练数据: 视频标题看成doc  
word: 字

char2vec

训练数据: 用户的观影看成doc  
word: 视频  
特征: 字向量

Topic2vec

训练数据: 用户的观影看成doc  
word: 视频  
特征: Topic+Topic embedding



字 embedding

Topic embedding

基于tf-idf对字向量加权

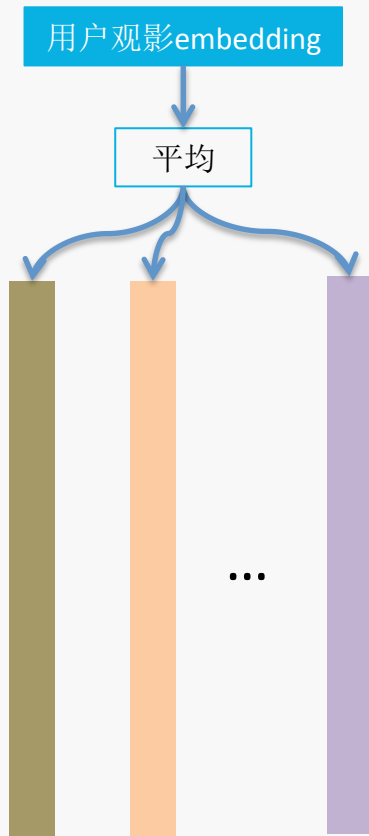
基于Topic加权

视频embedding

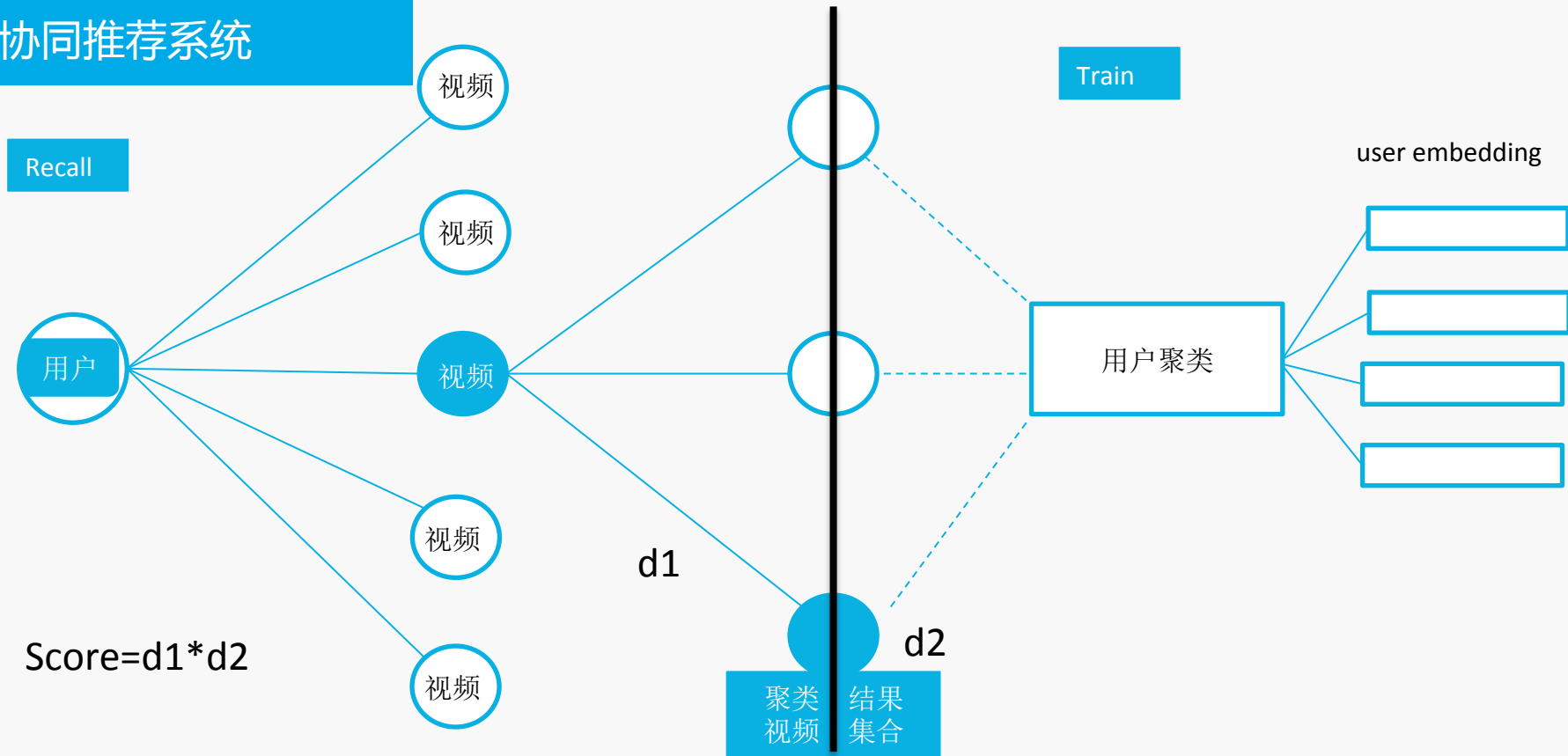
# 协同推荐系统

## UserCluster

- 输入
  1. 活跃用户观影序列
  2. 视频embedding(字向量生成)
- 计算方式
  1. 根据用户观影序列做embedding平均,得到用户embedding后,使用kmeans进行用户聚类



# 协同推荐系统



## 召回策略分析

- 分类和聚类:乒乓球,起点
- Item2item: 沙子,快速改变





# ▶ 精准排序系统(CTR)

用户行为数据:

CTR基础数据

Feature			Label
视频特征	用户特征	场景特征	是否点击
Video_Vector<k>	User_Vector<n>	Scenes_Vector<m>	0/1
v1	u1	s1	0
v2	u1	s1	1
...	...		...

- 精细评估,特征最多;
- 拟合点击率、播放时长等多目标;
- 耗时最严重一步。

## ▶ 精准排序系统(CTR)

### 传统机器学习模型

LR

本地AUC最小，训练速度最快，但上线效果一般

LightGBM

本地AUC最大，上线效果不错

XGBoost

本地AUC和上线效果同LGB，但训练时间过长

FM

本地AUC较大，上线效果较好，训练速度较快

LGB+LR/FM

本地AUC和上线效果都一般

### 深度学习模型

Wide&Deep

本地AUC和线上效果一般，正在持续优化

DeepFM

本地AUC与FM相差不大，但上线效果较FM好

NFM

本地AUC与FM相差不大，但上线效果较FM好

AFM和其它深度学习模型  
(开发中)

- 本地 AUC: LGB> NFM > DeepFM= FM >其他;
- 训练时间: NFM = DeepFM>lightGBM>FM>LR;
- 预测时间: LR=FM > lightGBM> nfm> DeepFM



SACC

第十届中国系统架构师大会  
SYSTEM ARCHITECT CONFERENCE CHINA 2018



# 助力推荐系统成长



十年架构 成长之路

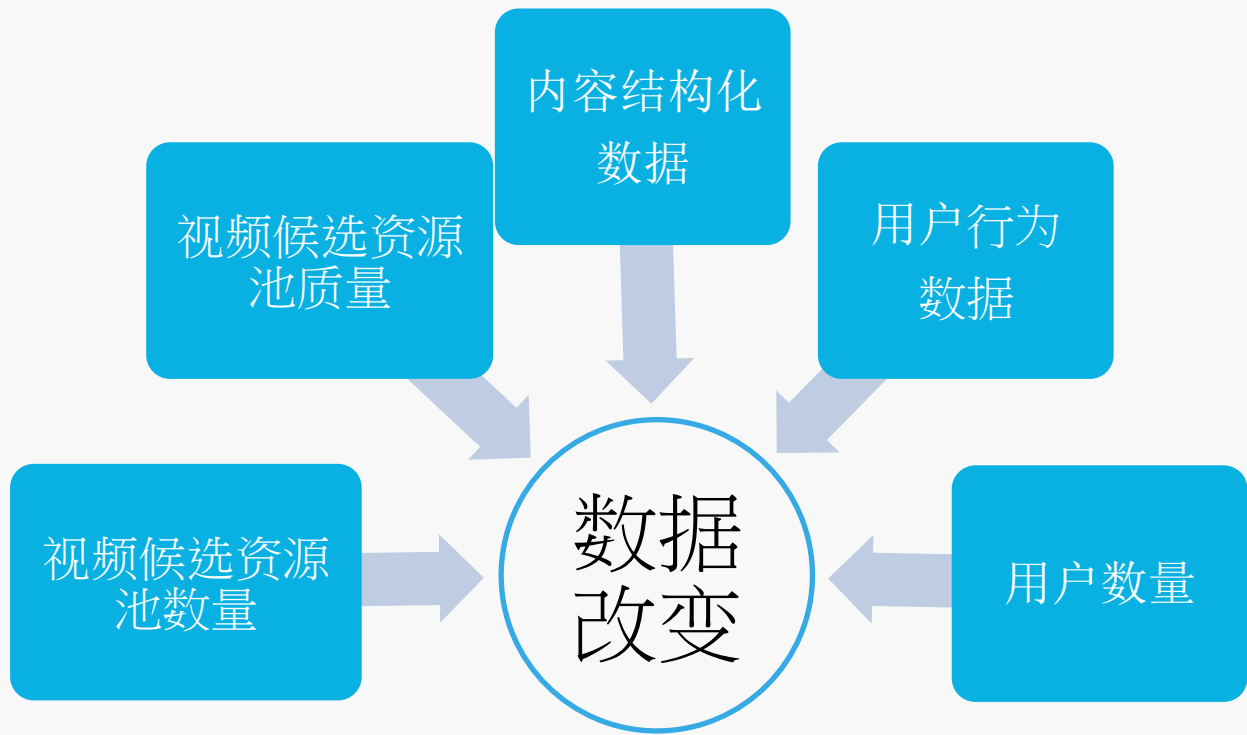


# 数据改变



十年架构 成长之路





# 数据改变

- 覆盖率高
- 可解释性强
- 可人工干预

内容结构化

行为数据

回顾:机器分发

热门排行榜

内容推荐

协同推荐

CTR

全局最热

聚合

item2item

Score<user,item>

推荐力度



粗

精

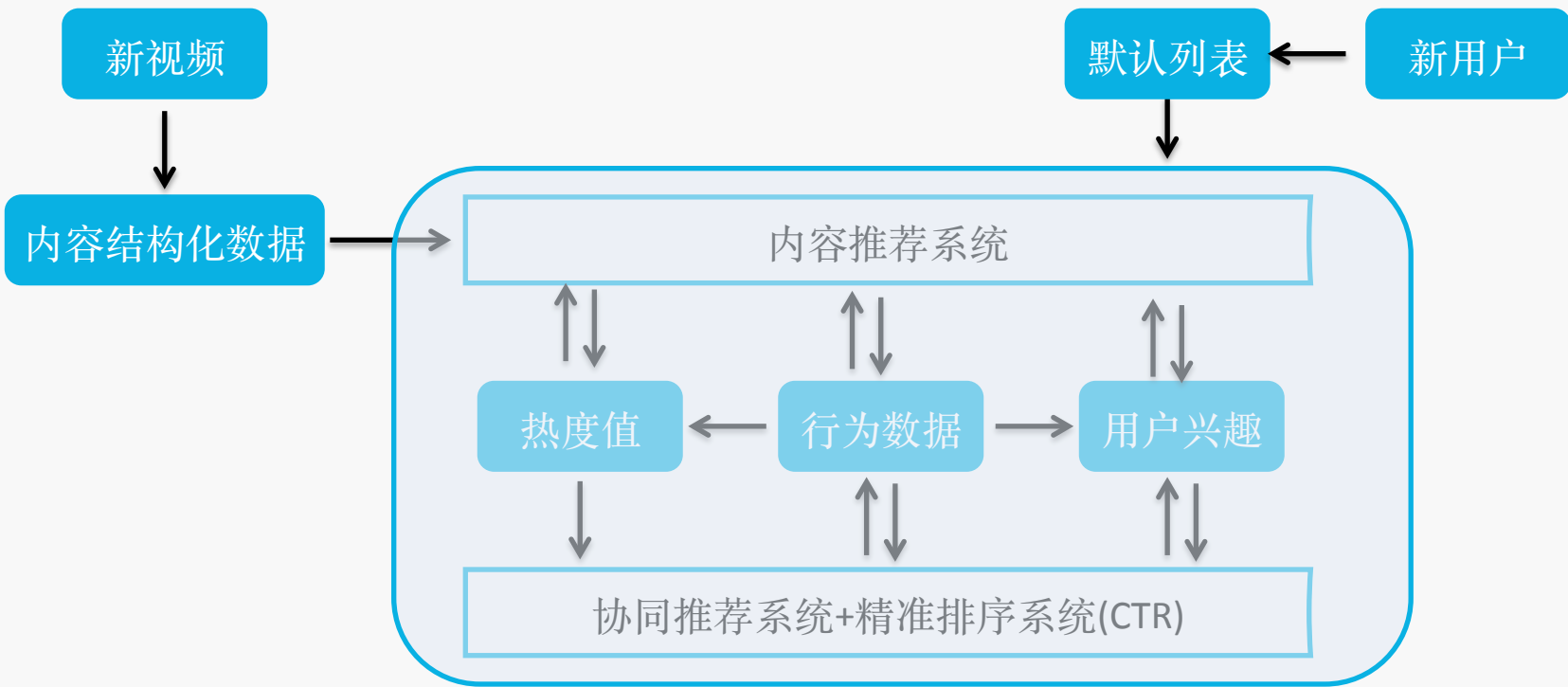
热度



高

低

## 数据改变



# 算法改变



十年架构 成长之路







- 视频结构化数据
- 用户行为数据

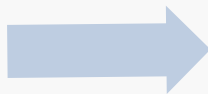


- 概率统计和预估
- 模型参数和特征
- 提升推荐细力度:降低热度值, 提升相关性

- 传统与DNN召回模型各有优势，DNN更有想象空间
- 行为数据作为目标度量、内容结构数据作为特征
- 点击率和覆盖率

## 深度学习模型

- FM系列深度学习模型
- 其它深度学习模型：跟进论文、自定义模型

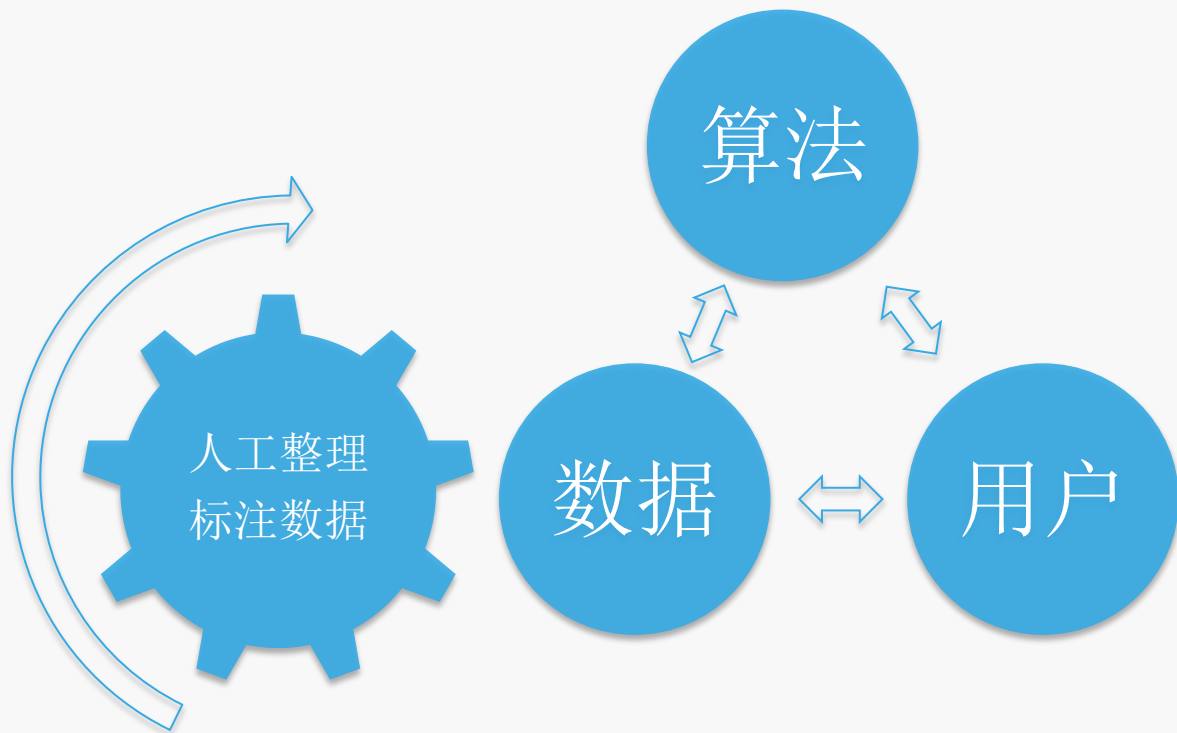


## 模型融合

分为相同输入 $x$ 和不同输入 $x$

- 传统机器学习模型互相融合
- 深度学习模型互相融合
- 深度学习模型与传统机器学习模型融合

## 总结



# 一些问题的思考 and 实践



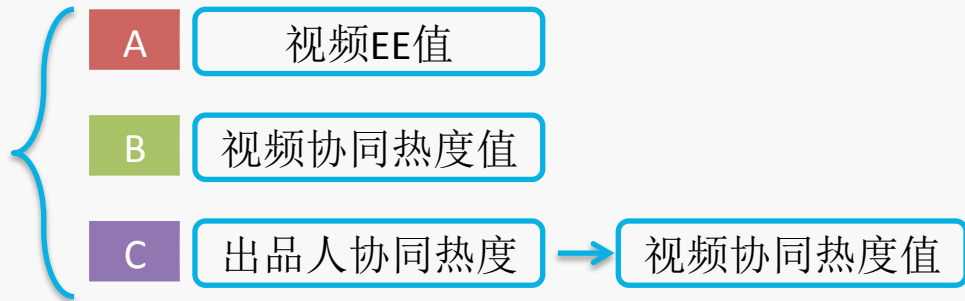
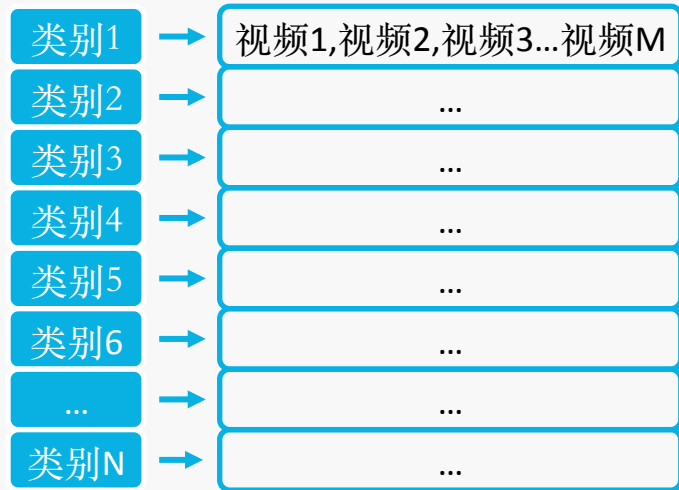
十年架构 成长之路



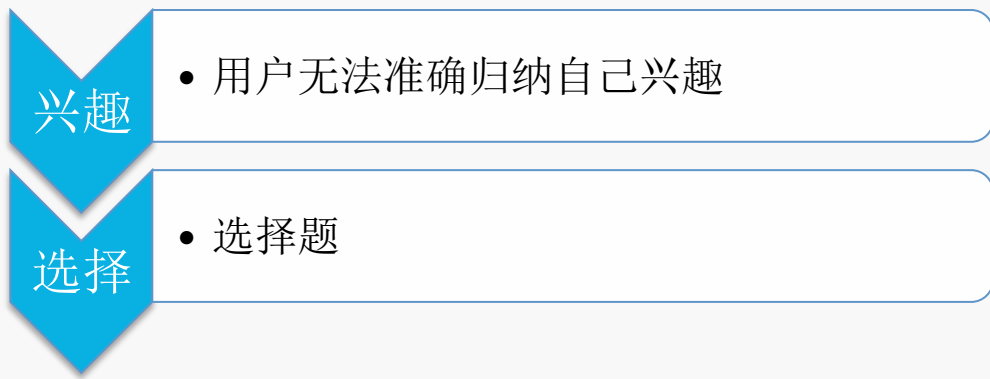
## ▶ 用户兴趣探索

默认列表:

- 最佳接入入口
- 保质量、保热度、保多样
- 类别和视频的**选择与排序**



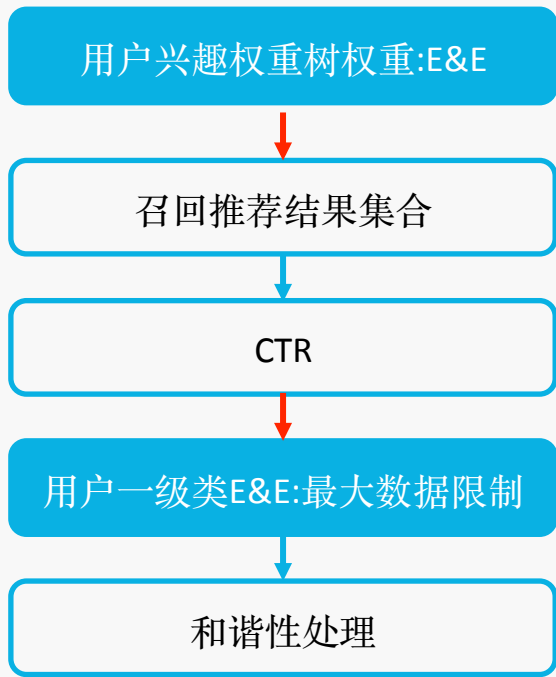
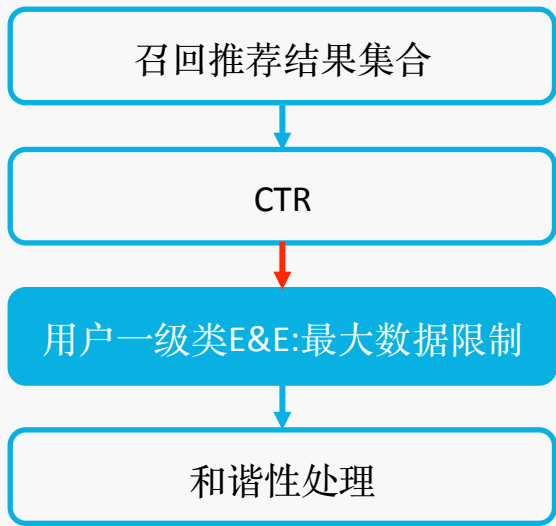
汽车大王亨利福特：“如果我当年去问顾客他们想要什么，他们肯定会告诉我，‘一匹更快的马’”



## 快速收敛、及时变化

用户画像增加负反馈

特征工程增加负反馈



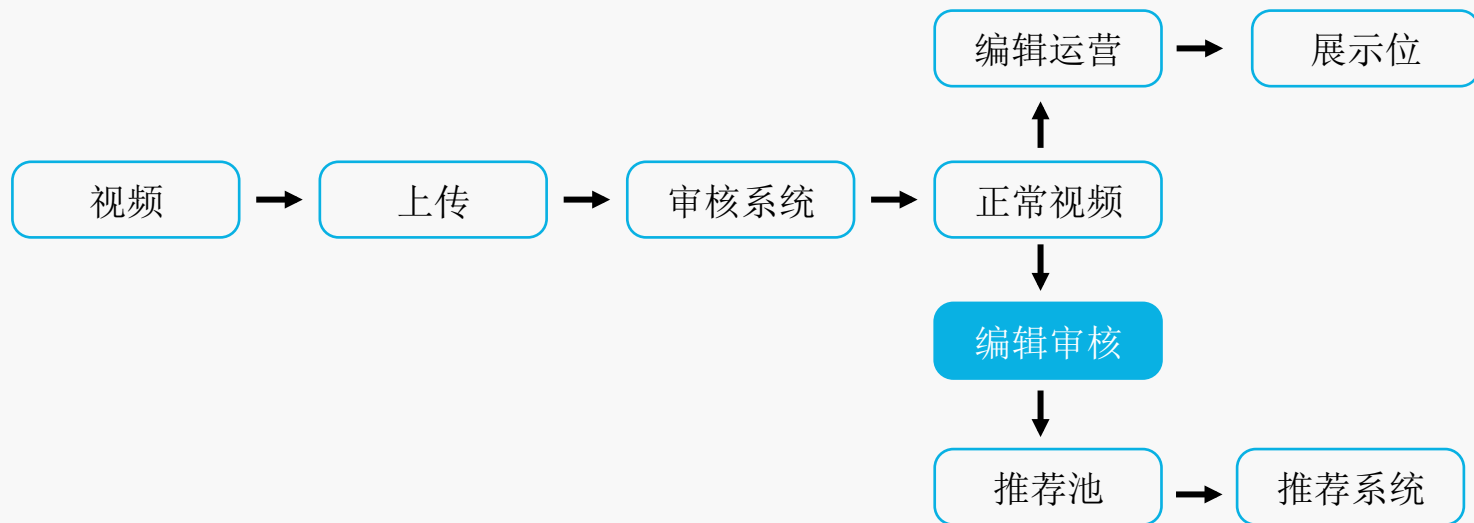


- 上层错误易继承:如内容分类
- 机器学习本身是有错误率，所以关键点需要人工参与，机器起到高效的辅助作用。
- 人工也存在主观错误率，所以多人评测，减少错误率。

Aesthetic Visual Analysis (AVA) 数据：

- 可靠性:每一张图有210个投票
- 普适性:括了专业的图像工作者，摄影师，也包括了摄影爱好者

## ▶ 内容质量



先验的内容质量标注，做优中选优更容易些

# 人工经验和用户消费等规律沉淀系统

“人工”整理和标注，把人类的思想认知、文化知识、心得经验带入视频画像中，通过内容推荐的探索方式产生行为数据，来弥补协同中的难以变化，给推荐系统提供进化的“源动力”



# THANKS

邮箱:[vocadata@foxmail.com](mailto:vocadata@foxmail.com)