



msup<sup>®</sup>

# 数据科学在音乐推荐中的 实践和应用

How Data Science is Boosting  
QQMusic Recommendation

<https://a2m.msup.com.cn/a2m2021/a2m2021/course?id=15492>





# 自我介绍



夏兵朝

QQ音乐数据科学中心

内容理解组负责人

从数据分析监控、数据BI工具建设以及专项数据挖掘，  
转型为闭环数据驱动内容宣发的数据科学，对内容评  
估理解和智能宣发有广泛深入的理解和实践





01

## QQ音乐内容分发 体系概览

- 1.1 场景介绍
- 1.2 系统架构介绍
- 1.3 问题定义和解决思路

02

## 数据科学基础 能力建设

- 2.1 指标体系建设
- 2.2 内容理解
- 2.3 用户理解

03

## 实践案例介绍1 - 冷启动

- 3.1 用户分层
- 3.2 用户冷启动
- 3.3 内容冷启动

04

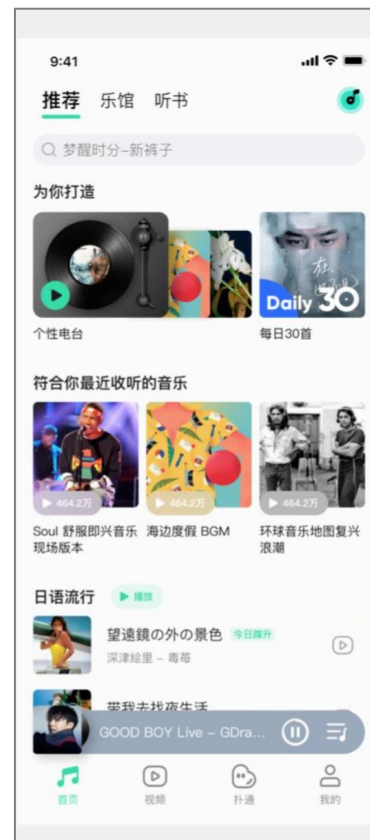
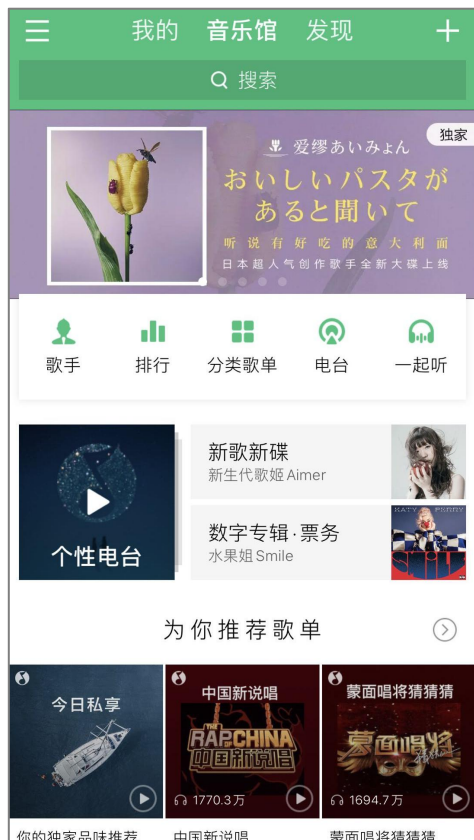
## 实践案例介绍2 - 内容生态和多样性

- 4.1 音乐场景的内容生态
- 4.2 用户体验的多样性
- 4.3 流量的利用和探索平衡  
问题





# 1.1 QQ音乐推荐场景介绍



8.0版本  
个性化功能上线  
感知体现



9.0版本  
独立页面分发  
场景独立

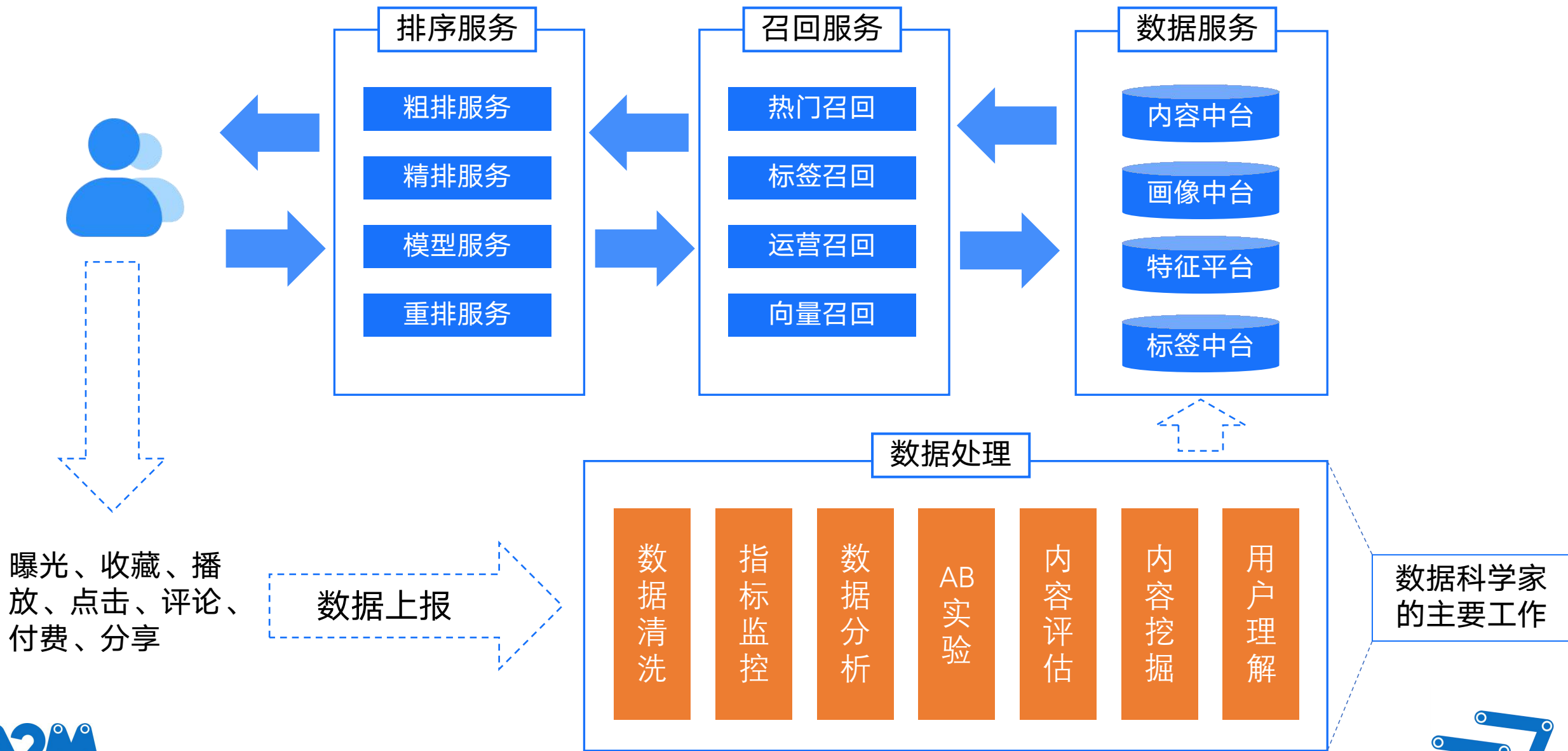


10.0版本  
全面个性化  
推荐功能整体渗透



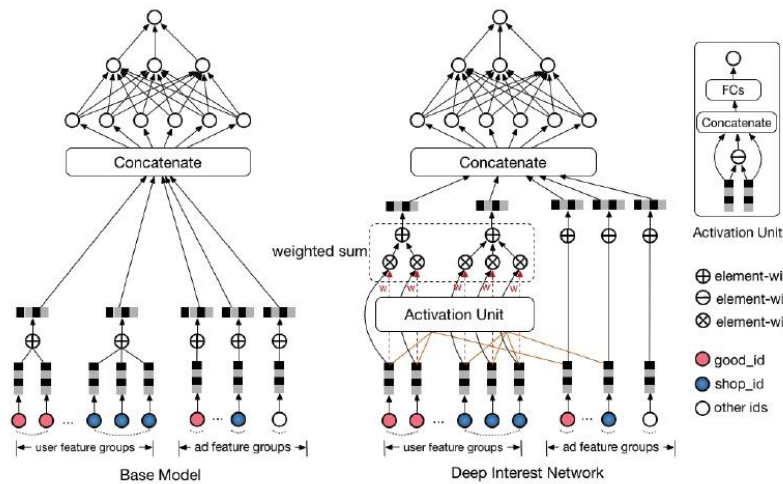
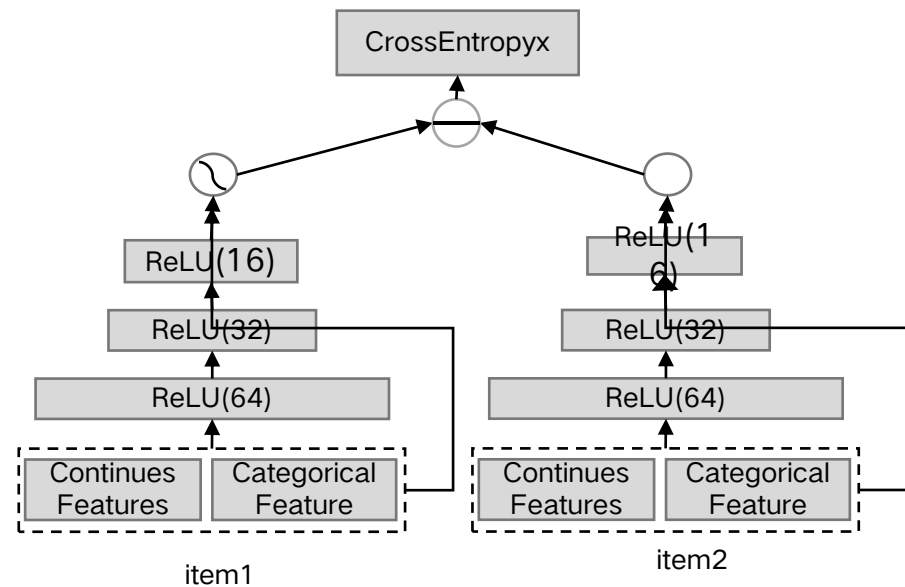
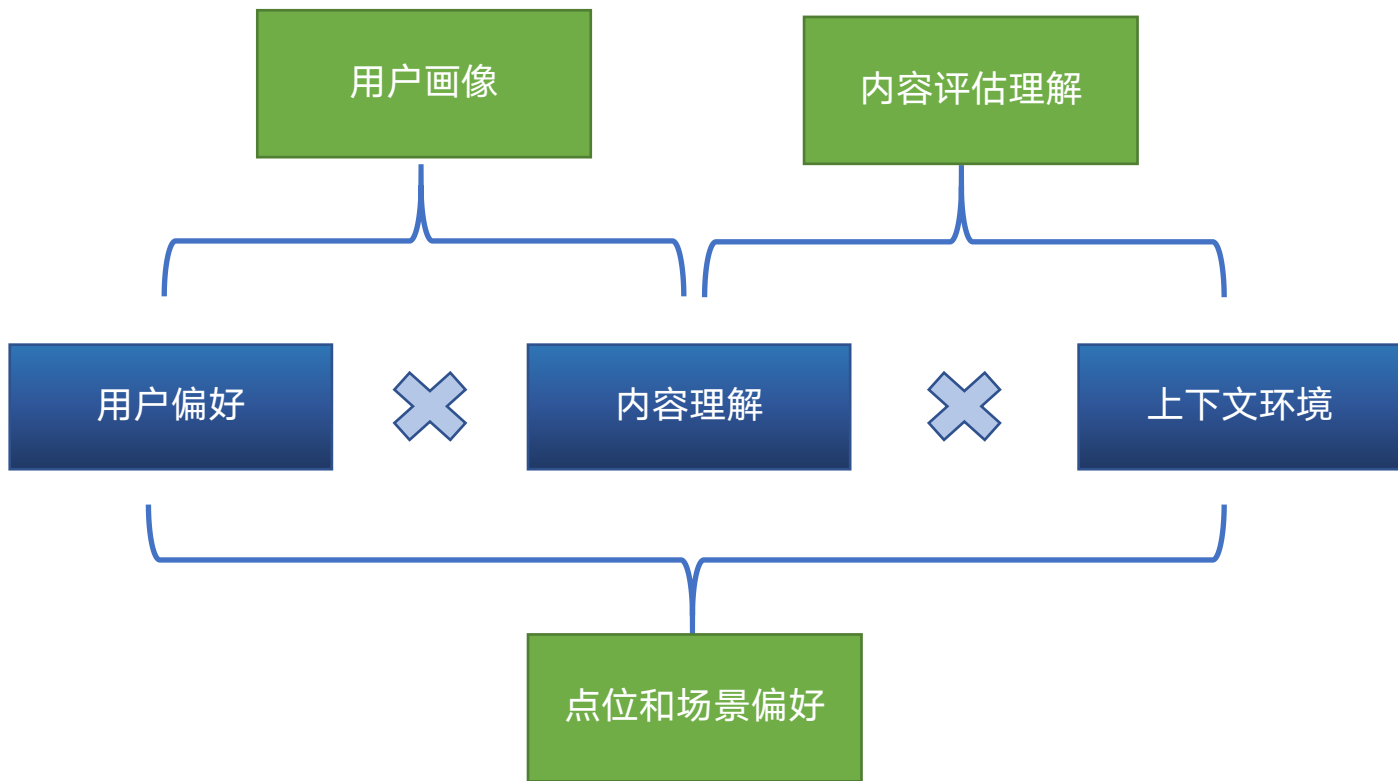


## 1.2 系统架构介绍





## 1.3 问题定义和解决思路





01

## QQ音乐内容分发 体系概览

- 1.1 场景介绍
- 1.2 系统架构介绍
- 1.3 问题定义和解决思路

02

## 数据科学基础 能力建设

- 2.1 指标体系建设
- 2.2 内容理解
- 2.3 用户理解

03

## 实践案例介绍1 - 冷启动

- 3.1 用户分层
- 3.2 用户冷启动
- 3.3 内容冷启动

04

## 实践案例介绍2 - 内容生态和多样性

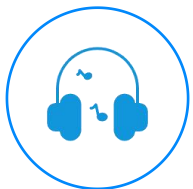
- 4.1 音乐场景的内容生态
- 4.2 用户体验的多样性
- 4.3 流量的利用和探索平衡  
问题







## 2.1 指标体系建设



用户消费

- 听歌渗透 (DAU)
- 用户时长 (总/人均时长)
- 功能留存 (次留/周留)
- 互动渗透 (点赞率/收藏率)



内容生产

- 内容生产 (新/热内容)
- 内容准入 (分渠道)
- 内容池生态 (分标签/垂类扶持)



推荐分发

- 分发效率 (热点/探索)
- 内容多样性/覆盖率 (基尼系数)
- 服务稳定性 (召回/排序, 特征服务)







## 2.2 内容理解概述

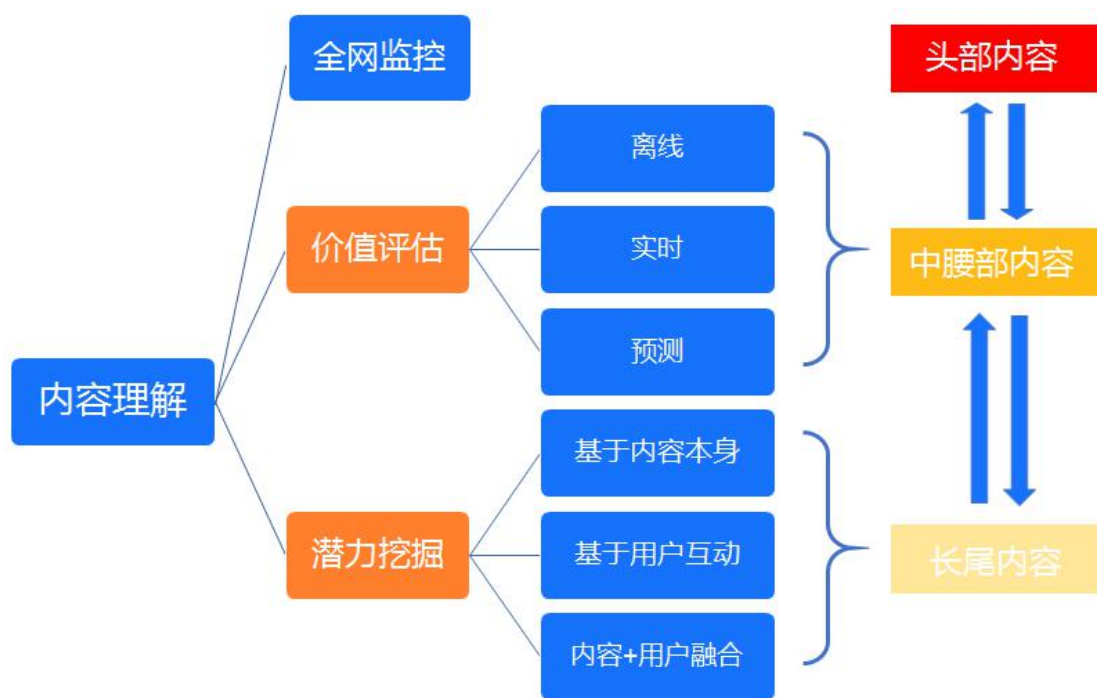
### ■背景：内容分发的“长尾分布”

- 20%的头部内容占据了80%以上的播放量
- 长尾内容很难获取到真实可信的用户反馈数据
- 不断有新增的内容上架



### ■解决方案：“分而治之”

- 中头部内容利用丰富的用户反馈数据进行价值评估
- 长尾内容通过稀疏的用户反馈数据进行挖掘和探索
- 冷启内容通过内容本身来理解和挖掘





# 2.2.1 内容离线评估方案

■ 基础评估维度

- 用户互动质量评估
- 付费驱动能力评估
- 用户圈层指数
- 拉新拉活能力
- 飙升趋势预测
- 实时数据评估

■ 业务策略维度

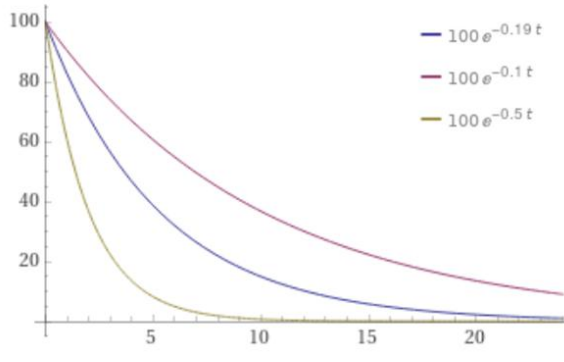
- 集团优选
- 腾讯音乐人
- 重点流派
- 付费内容



$$\hat{R} = \frac{C + \alpha}{I + \alpha + \beta}$$
$$\alpha = \left( \frac{\bar{R}(1 - \bar{R})}{S^2} - 1 \right) \bar{R}$$
$$\beta = \left( \frac{\bar{R}(1 - \bar{R})}{S^2} - 1 \right) (1 - \bar{R})$$

其中  $\bar{R}$ 、 $S^2$  分别为 xx 率的均值、方差

贝叶斯平滑

$$T(t) = T(t_0) * e^{-k(t_0-t)}$$


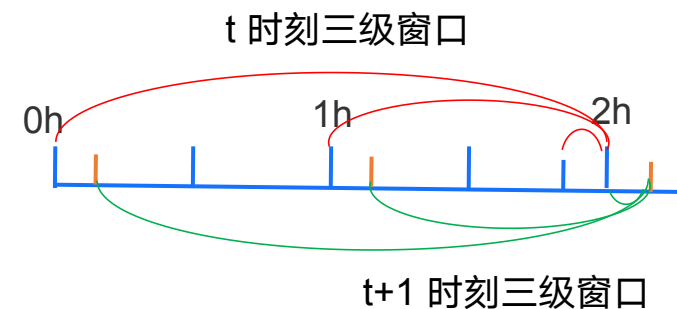
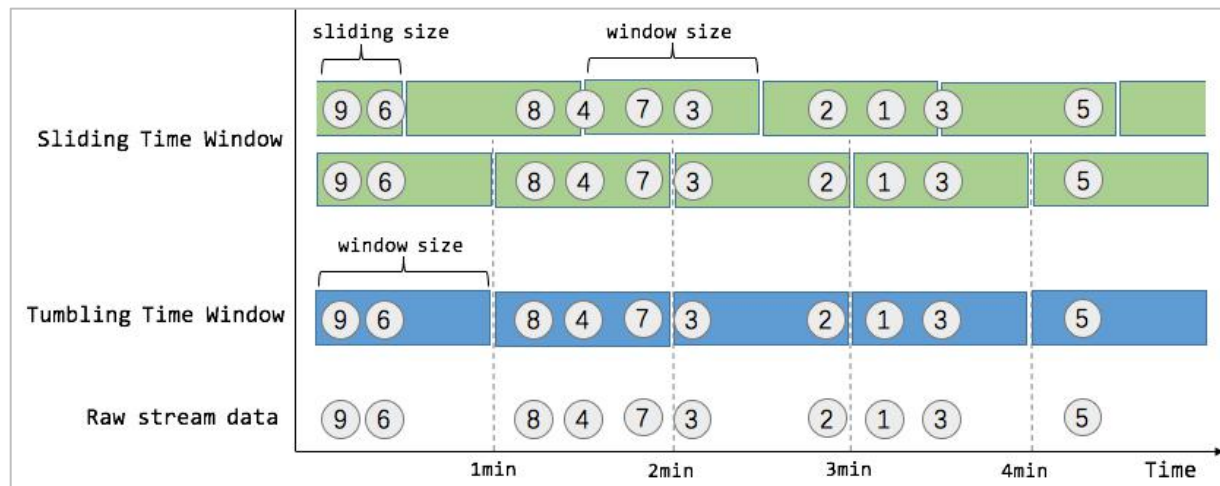
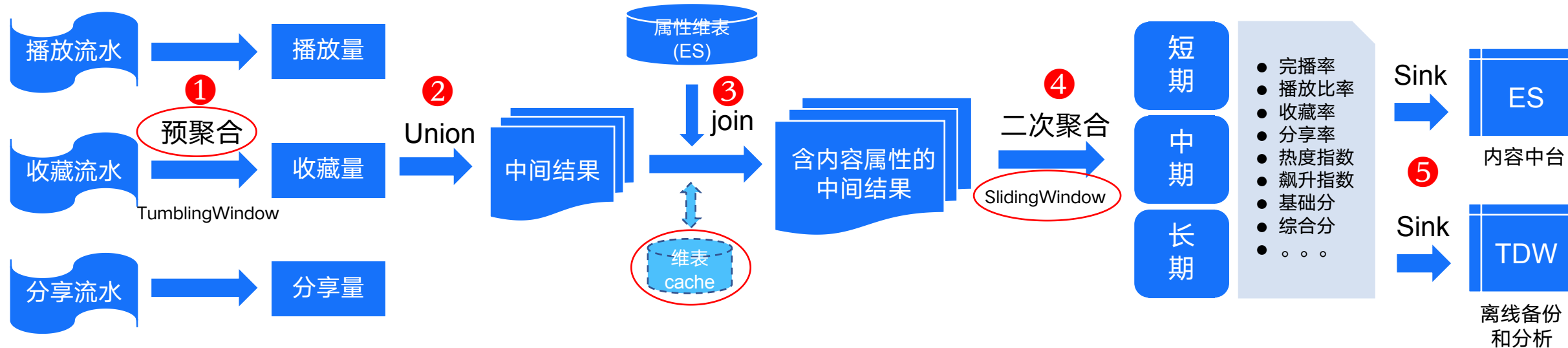
牛顿冷却定律





## 2.2.2 内容实时评估方案

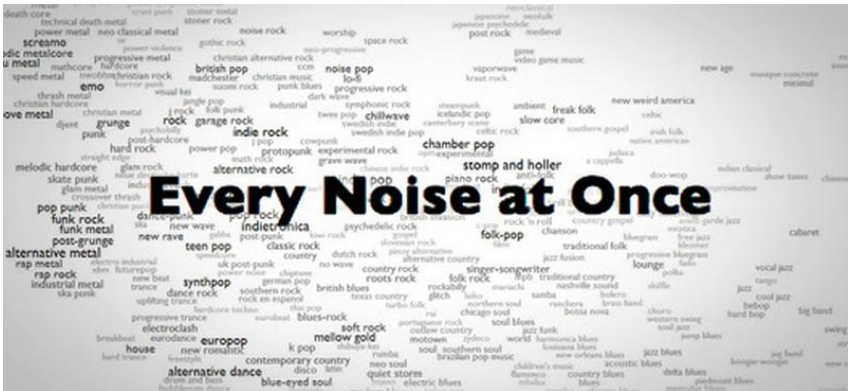
技术方案： 基于Apache Flink进行流式数据计算





# 2.2.3 内容的嵌入表征

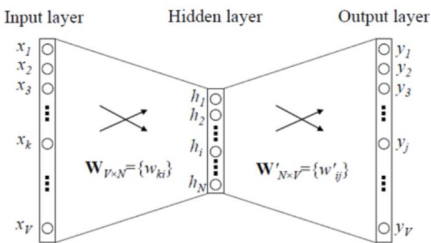
内容语义特征：  
基于内容固有属性和标签



Step1 语料选择

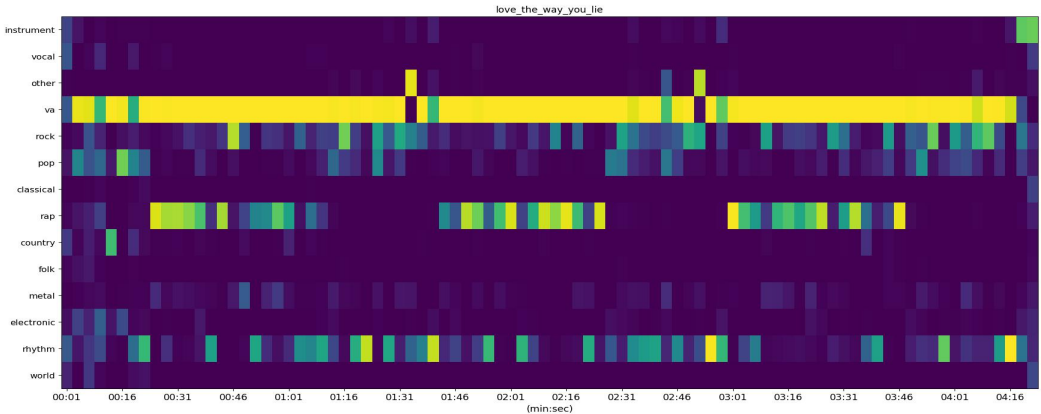
User-A	Song1	Song2	Song3	
User-B	Song2	Song3	Song5	...
User-C	Song3	Song5	Song6	...
User-D	Song3	Song9	Song6	...

Step2 模型训练



内容相关性：  
基于用户行为进行计算

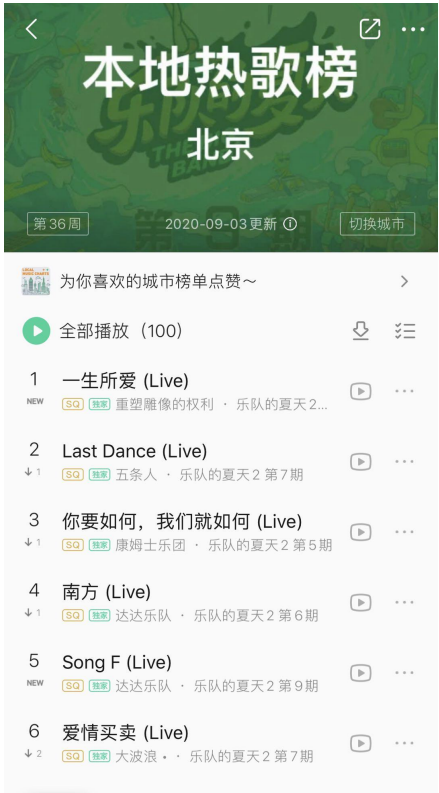
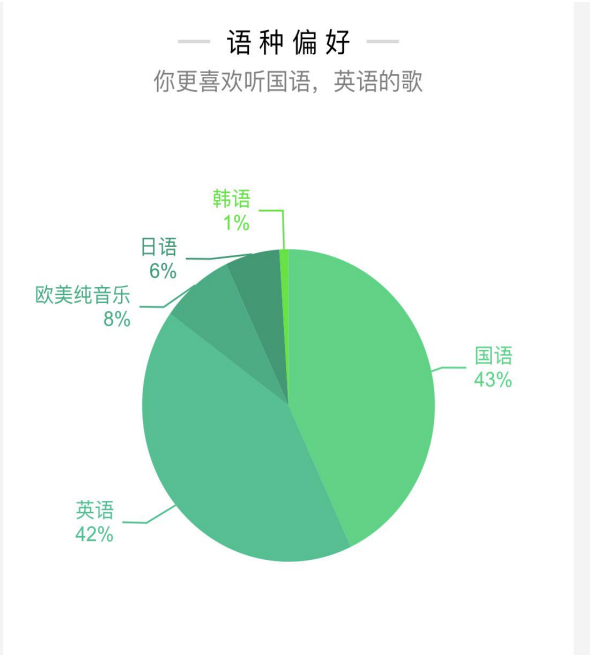
内容音频特征：  
基于音乐本身量化





# 2.3 用户理解

- 用户行为 X 内容理解 = 用户喜好画像
  - 实时更新
  - 多种类：流派，语种，歌手，歌曲
- 用户固有属性 X 内容消费 = 圈层偏好
  - 周期更新
  - 多层次：年龄，性别，城市等级







01

## QQ音乐内容分发 体系概览

- 1.1 场景介绍
- 1.2 系统架构介绍
- 1.3 问题定义和解决思路

02

## 数据科学基础 能力建设

- 2.1 指标体系建设
- 2.2 内容理解
- 2.3 用户理解

03

## 实践案例介绍1 - 冷启动

- 3.1 用户分层
- 3.2 用户冷启动
- 3.3 内容冷启动

04

## 实践案例介绍2 - 内容生态和多样性

- 4.1 音乐场景的内容生态
- 4.2 用户体验的多样性
- 4.3 流量的利用和探索平衡  
问题





# 3.1 用户分层

## 活跃分层

未登录

低资产

核心活跃

## 垂类分层

口碑用户

校园圈层

年轻人群体

## 画像分层

探索型

多元型

新歌型

冷门型

独特型

## 独爱小众新口味

Biubiu的音乐偏好



**探索**  
★★★

你就像一名经验丰富但从不满足的探险家，总在试探和品析那些未曾听过的旋律

**专一**  
★★

你热衷于最爱的音乐旋律，且相信那是上天给予自己最好的礼物

**独特**  
★★

有人会说你的品味独特，但也可能是你有着超前的音乐审美，发现大家不易发现的宝藏

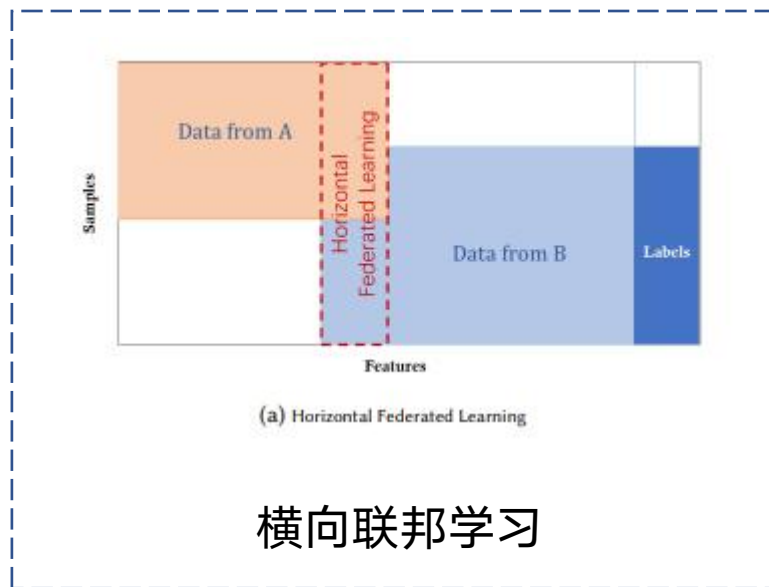




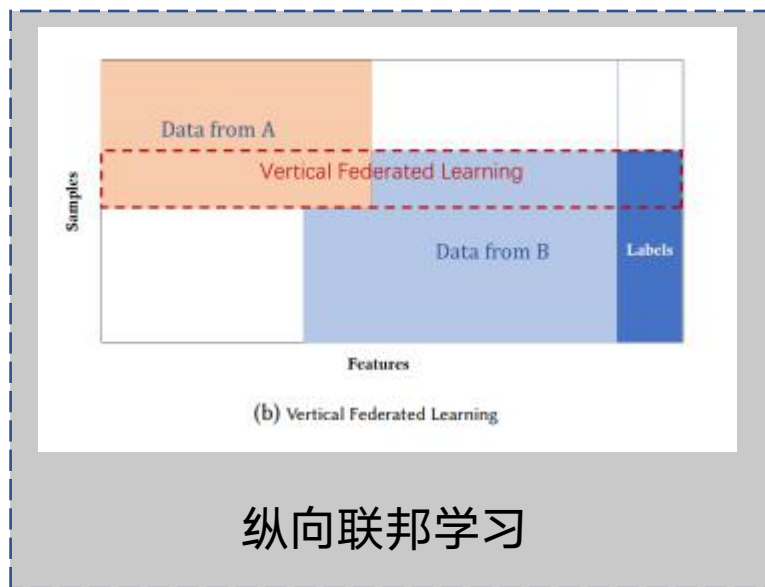


## 3.2 用户冷启动——纵向联邦学习

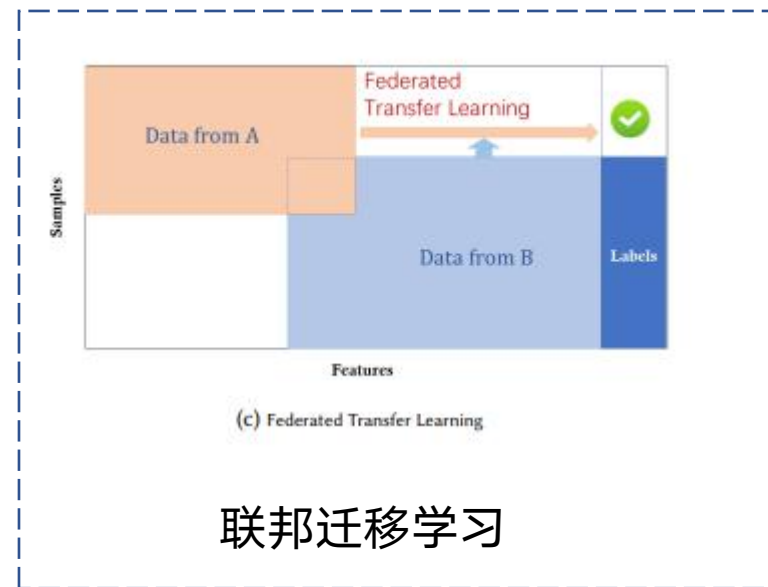
联邦学习是一种机器学习技术，可在拥有本地数据样本的多个分布式边缘设备或服务器之间训练算法，而无需交换数据样本，保护数据隐私。近年随着联邦学习的兴起，在金融等领域已经有多个联合建模的成功案例，我们也开始寻求在大腾讯生态下引入**纵向联邦学习**提升召回的准确性



- 业务相同或相似
- 特征重叠多
- 样本联合



- 触达用户相似
- 用户重叠多
- 特征联合

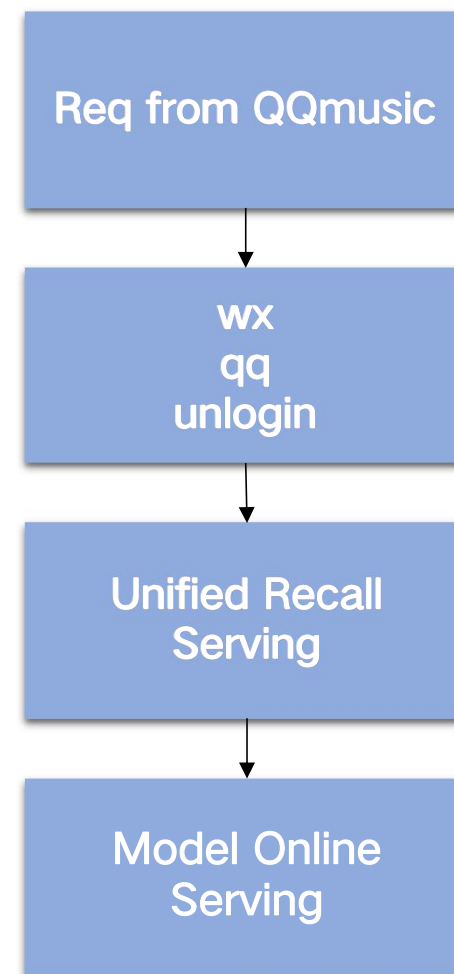
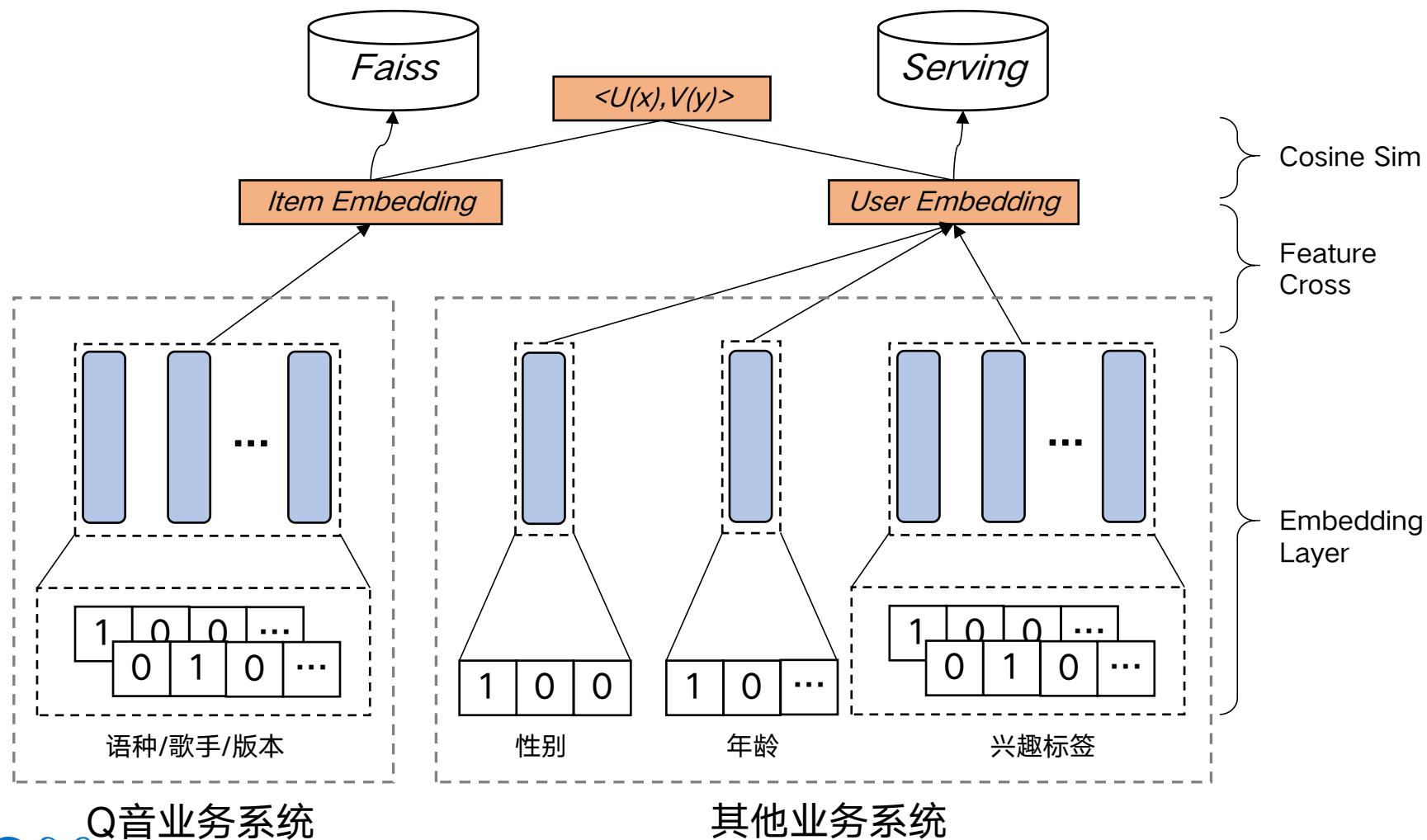


- 用户业务均不相似
- 特征和用户重叠少
- 迁移学习



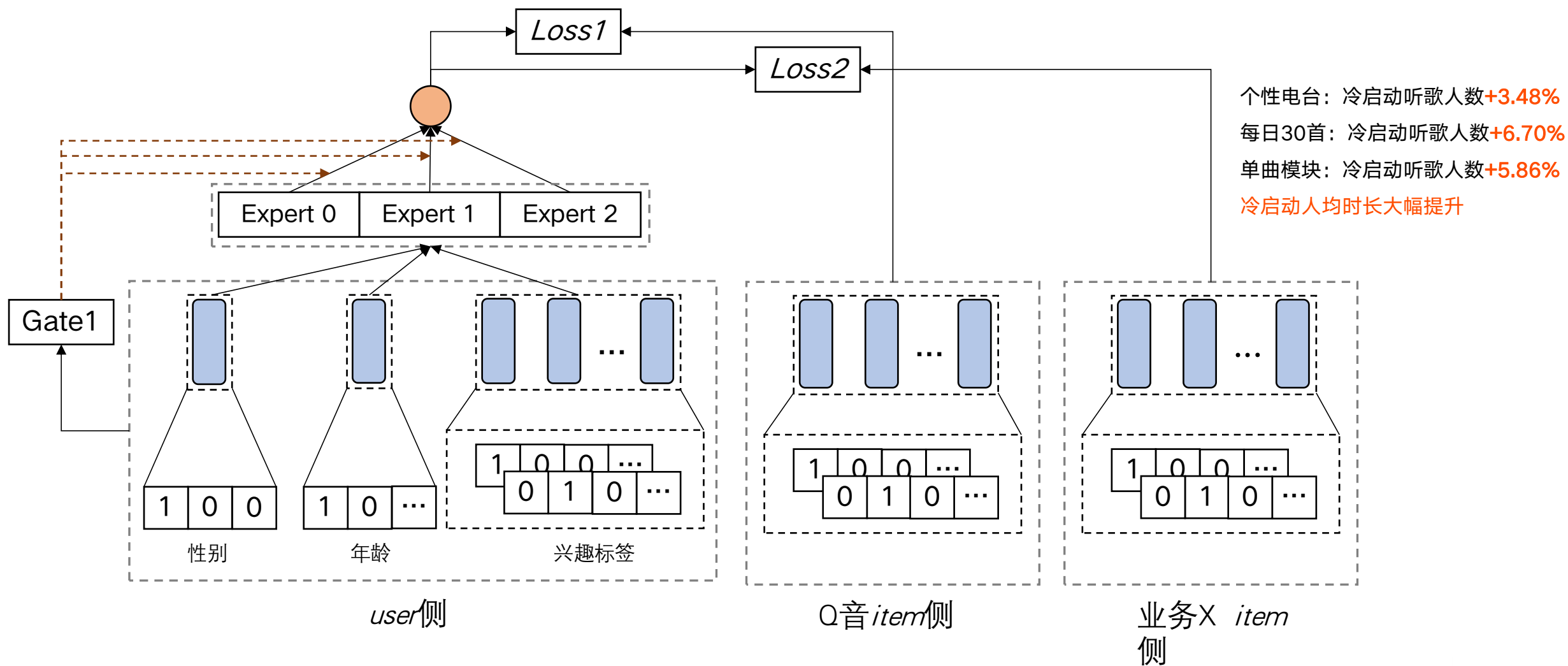


## 3.2 用户冷启动——纵向联邦学习-DSSM





## 3.2 用户冷启动——纵向联邦学习-MMoE多目标



注: TME严格遵守相关法律法规, 遵循隐私保护原则, 为用户提供更加安全、可靠的服务





# 3.3 内容冷启动

音乐本身包含非常多的固有属性，例如专辑、歌手等等，为了提升召回的准确性，很多召回模型会将其作为歌曲的Side-Info融合进模型进行学习，在QQ音乐召回中我们使用了EGES/GraphSage



歌曲的Meta通常作为模型特征输入



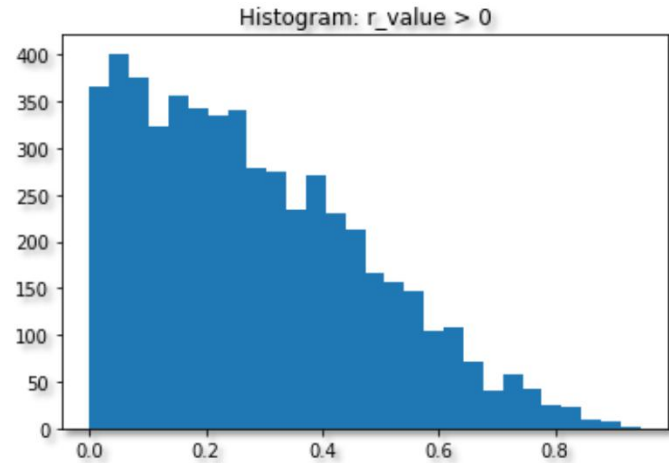
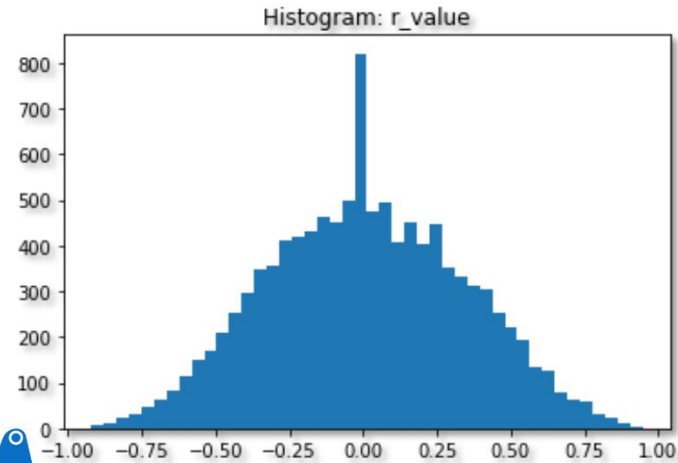
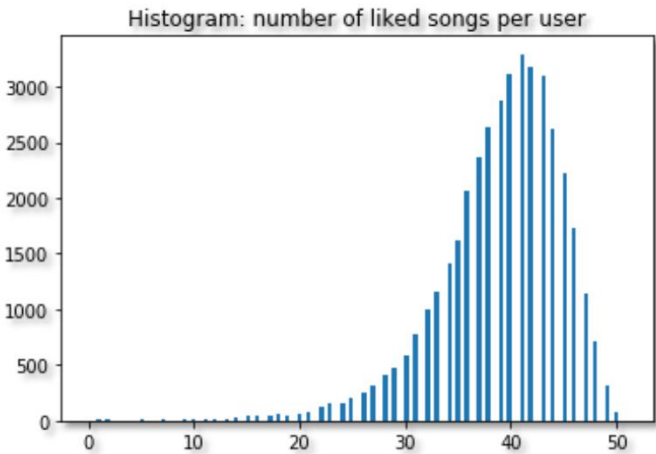
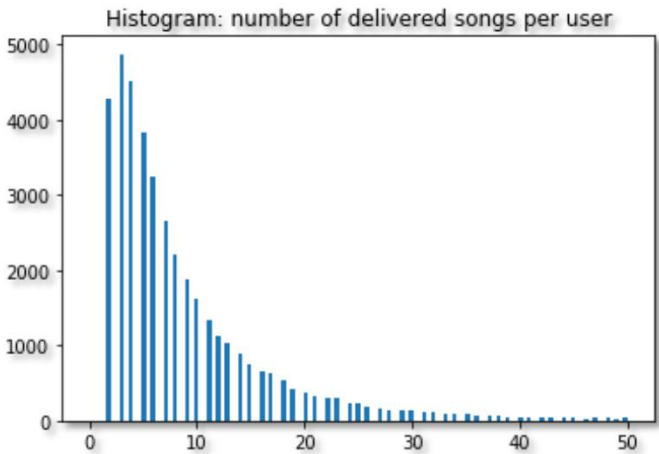
EGES融合Side-Info	GraphSage中作为节点特征
歌手/语种等特征的增加让召回的泛化性变差，DY生态洞穿了很多歌曲的Meta关联逻辑	歌手/语种/歌词等特征作为图节点特征融入图模型提升召回的准确性





# 3.3 内容冷启动

我们发现歌曲与用户资产的音频embedding加权相似度与用户听歌完播率的相关系数符合正态分布，从某种程度上说明部分用户听歌行为与音频是敏感的( $r\_value > 0$ )



用户	歌曲	完播率	加权音频相似度
$u(1)$	$s[m, u(1)]$	$p[m, u(1)]$	$S[m, u(1)]$
...	...	...	...
$u(i)$	$s[m, u(i)]$	$p[m, u(i)]$	$S[m, u(i)]$

计算皮尔逊相关系数

以用户第 $m$ 首收听的歌曲为例，其embedding表示为 $d(m)$ ， $d(m)$ 和该用户 $N$ 首资产相应的音频embedding即 $l(n)$ ,  $n=1,2,...,N$ 的加权相似度 $S(m)$ 为:

$$S[m, u(1)] = \frac{\sum_n w(n) * \cos\_sim(l(n), d(m))}{N}$$





# 3.3 内容冷启动

结合分析结论，音频的embedding用在了Q音单曲推荐多个场景的召回模块中，在没有其他协同信息的情况下，挖掘歌曲音频表征也有助于冷启动分发

## 基于音频相似召回



User

单点召回



Song

### 基于音频表征召回歌曲

Camila Cabello - Havana

google-choise - Different

World

Justin Bieber - Peaches

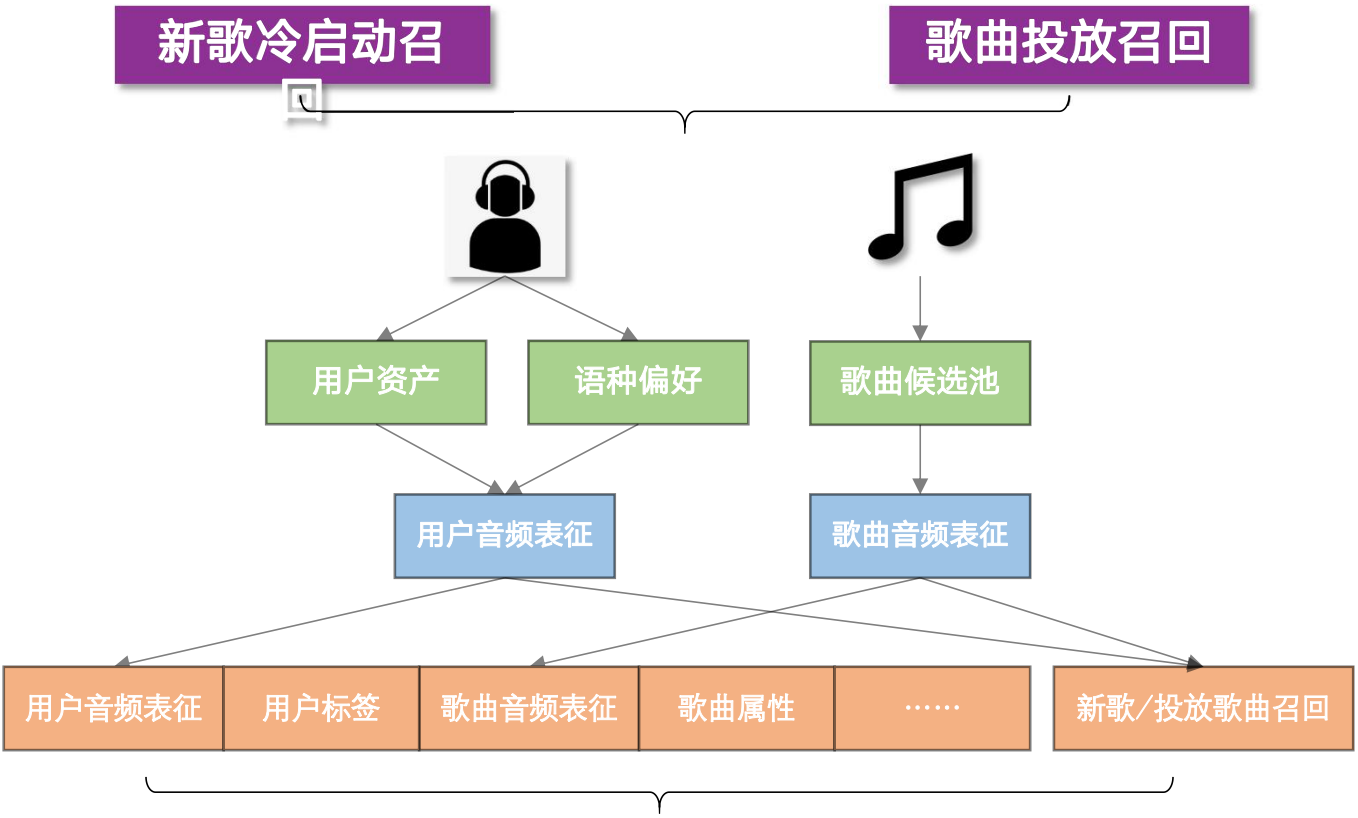
Troye Sivan - For him.

G.E.M. 邓紫棋 - Walk on Water

### 用户偏好歌曲

Leave The Door  
Open

(Bruno Mars 大热单  
曲)



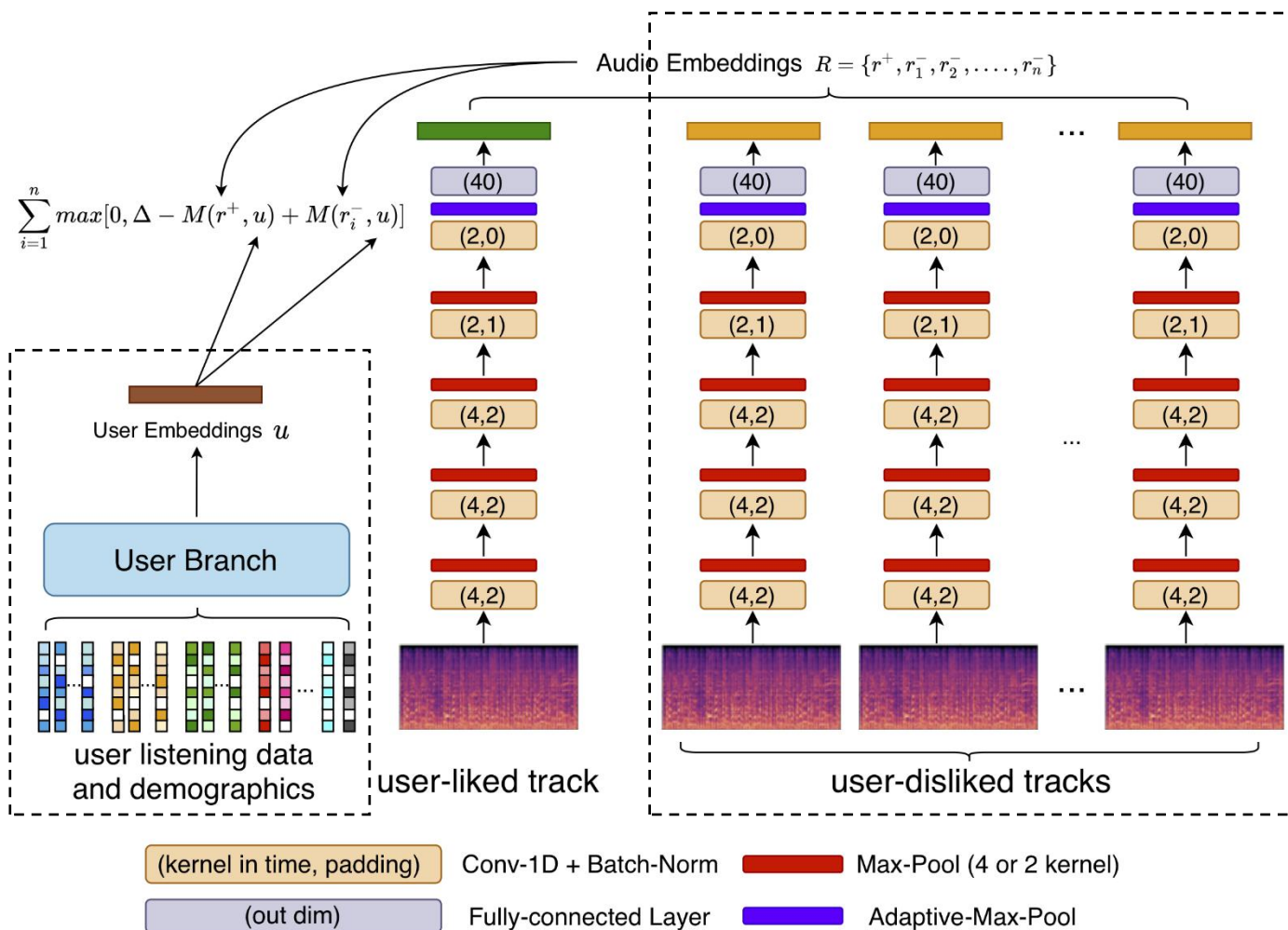
新歌/投放歌曲DeepFM和DCN排序







## 3.3 内容冷启动



- ✓ User部分的模型, 采用深度模型计算40维U
- ✓ Audio部分的模型, 改用用户喜欢的1首歌vs不喜欢的n首歌, 与40维user embedding做metric learning
- ✓ 训练好audio部分的模型, 可对任何音频输入得出embedding 也是40维
- ✓ 相比于前面提到的audio embedding, 融合了user信息的user audio embedding在音频召回的准确率上得到了进一步提升; 这一点也在MIREX大奖中country, rap/hip-hop/K-pop这3个流派分类精确度达到历史最好成绩

User-Audio Embedding模型拿下MIREX大奖, 论文发表在ICASSP

\*Learning Audio Embeddings with User Listening Data for Content-based Music Recommendation." (ICASSP), 2021







01

## QQ音乐内容分发 体系概览

- 1.1 场景介绍
- 1.2 系统架构介绍
- 1.3 问题定义和解决思路

02

## 数据科学基础 能力建设

- 2.1 指标体系建设
- 2.2 内容理解
- 2.3 用户理解

03

## 实践案例介绍1 - 冷启动

- 3.1 用户分层
- 3.2 用户冷启动
- 3.3 内容冷启动

04

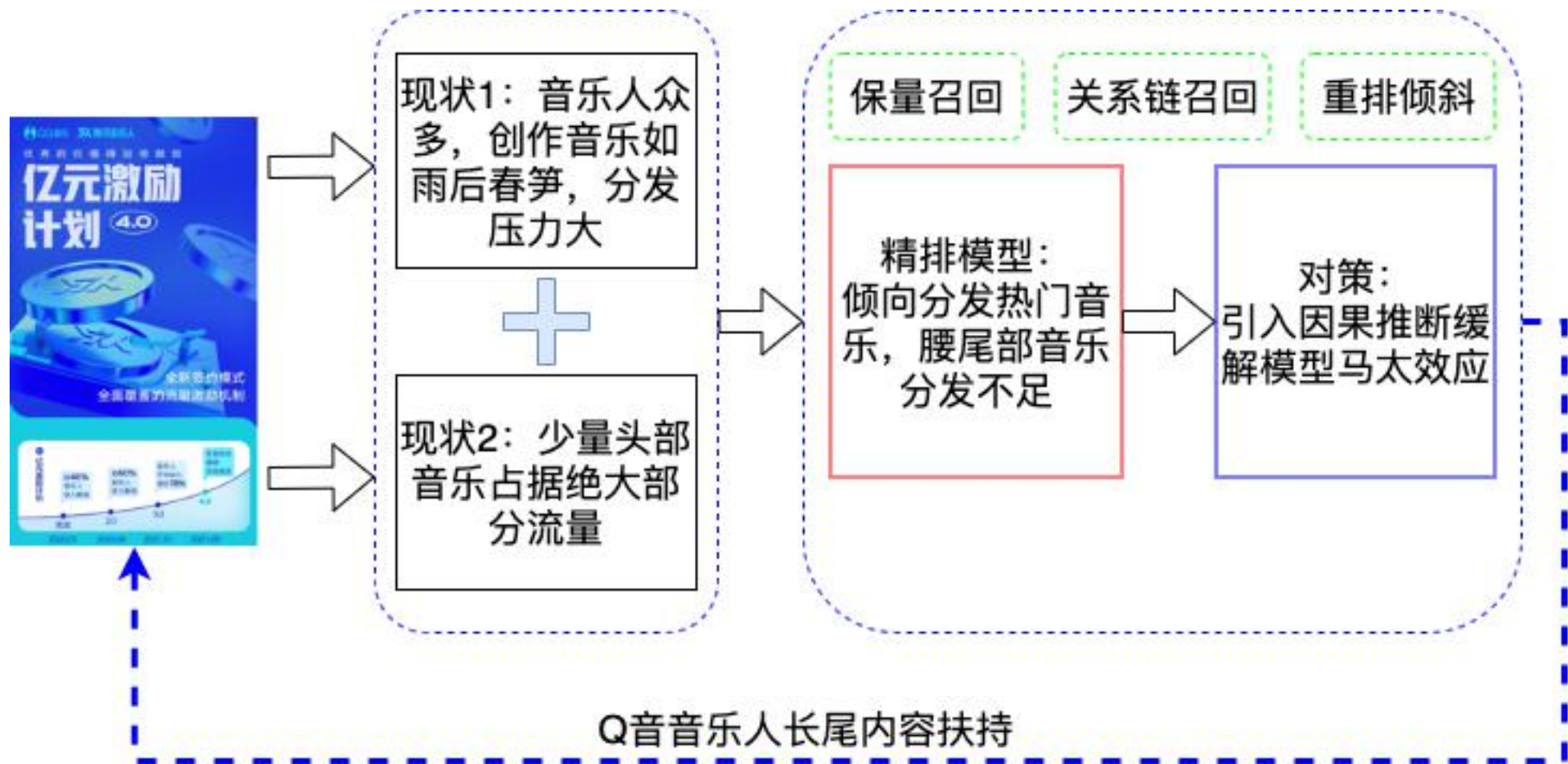
## 实践案例介绍2 - 内容生态和多样性

- 4.1 音乐场景的内容生态
- 4.2 用户体验的多样性
- 4.3 流量的利用和探索平衡  
问题





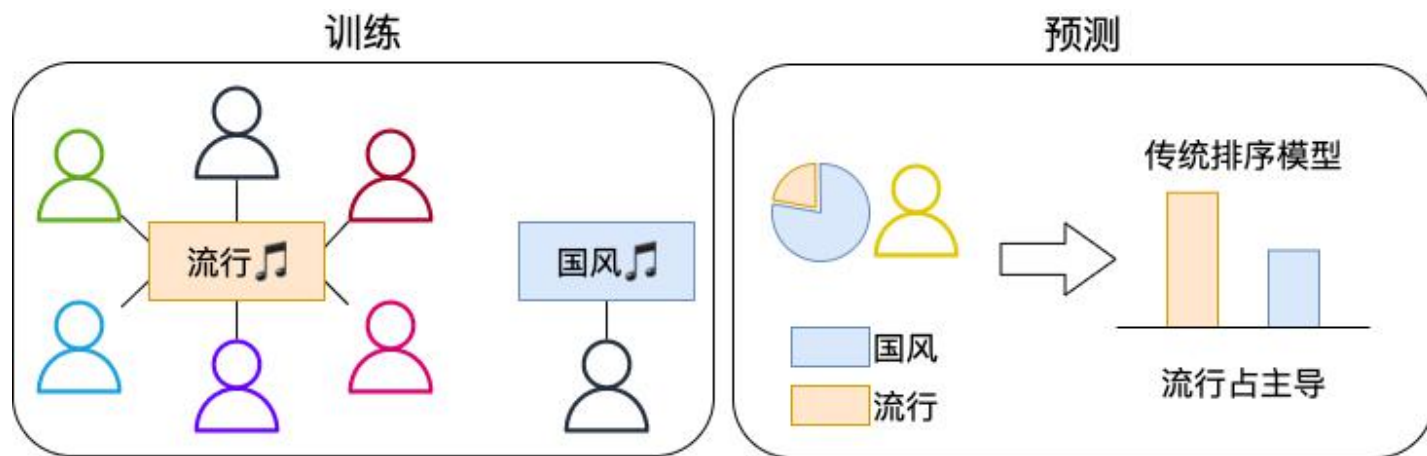
## 4.1 音乐场景的内容生态



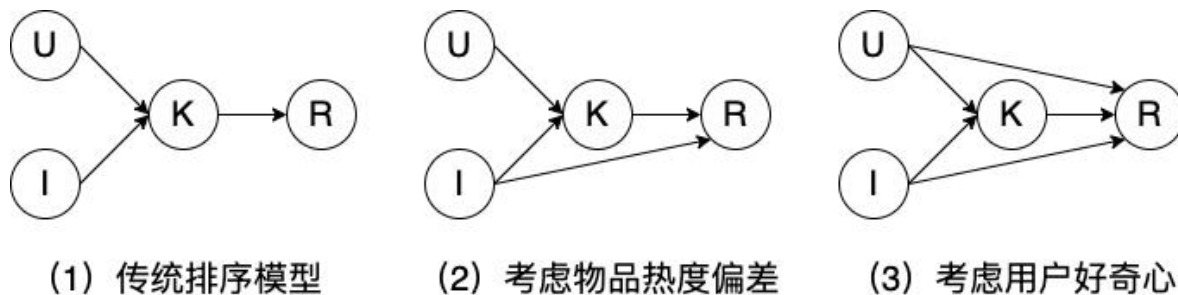


## 4.1 音乐场景的内容生态

**问题：** 传统精排模型存在马太效应（头部效应）。



**方案：** 引入因果推断考虑物品热度偏差和用户好奇心带来的点击率影响，学习用户对物品真实偏好





## 4.1 音乐场景的内容生态

具体实现：

【训练过程】同时考虑三个因素对点击率的影响：

- 用户对物品真实偏好
- 用户好奇心
- 物品流行度

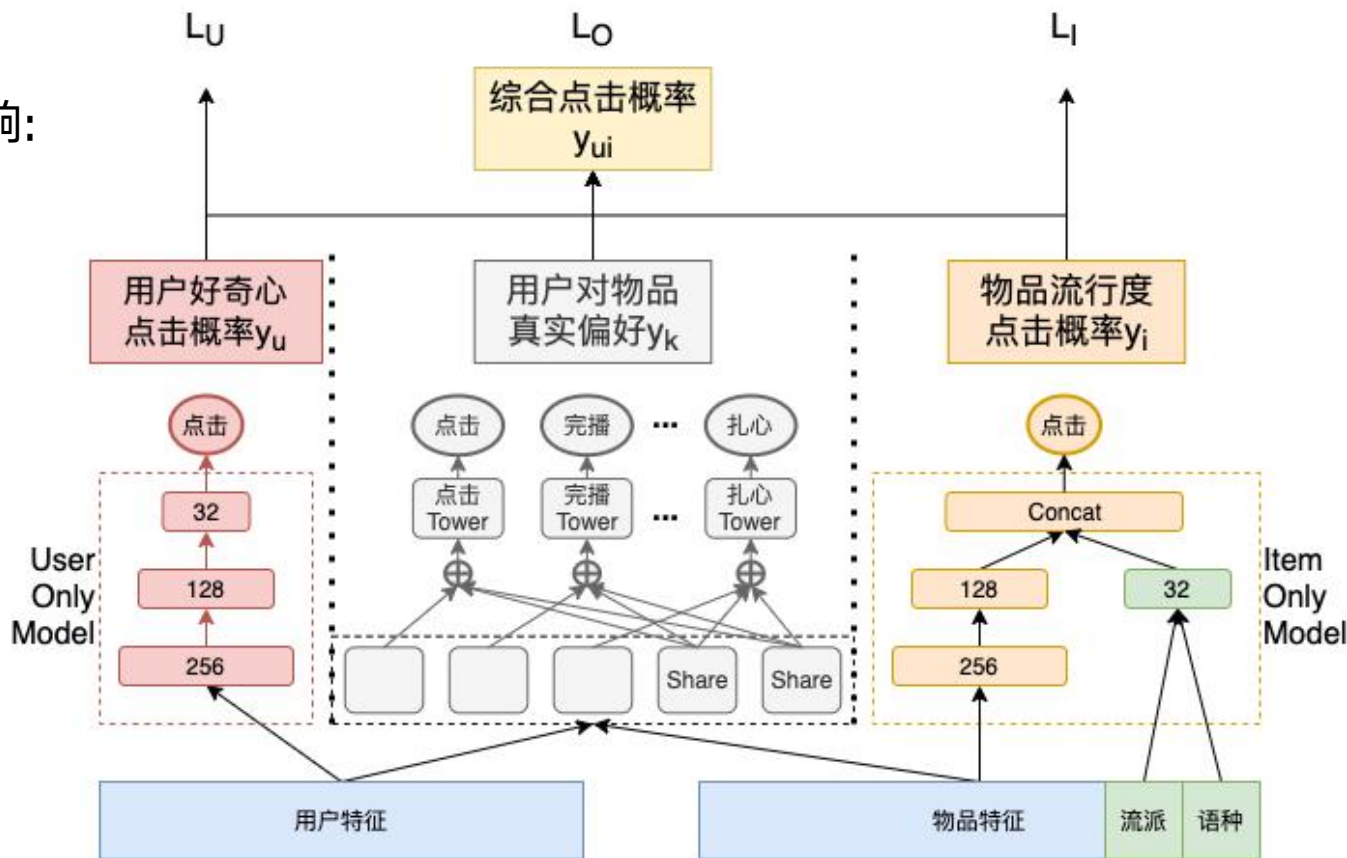
【预测过程】即可得到用户对物品真实偏好：

$$soe_{ui} = y_{ui} - c * y_u * y_i$$

c为超参数。

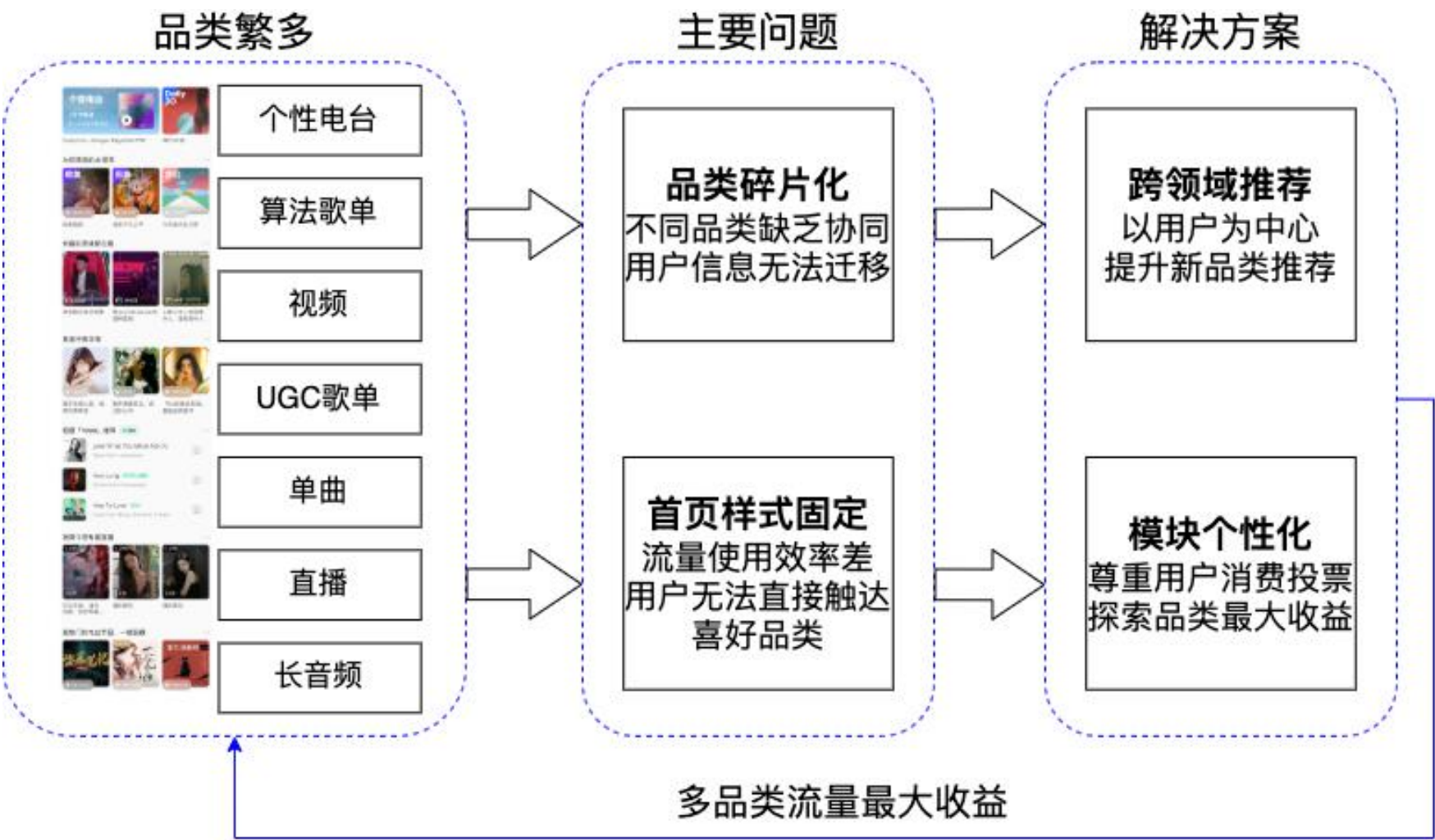
效果：

- 既缓解马太效应，扶持长尾内容——内容分发数量提升13%
- 又进一步理解用户偏好，优化推荐体验——用户收藏率提升3%





# 4.2 用户体验的多样性



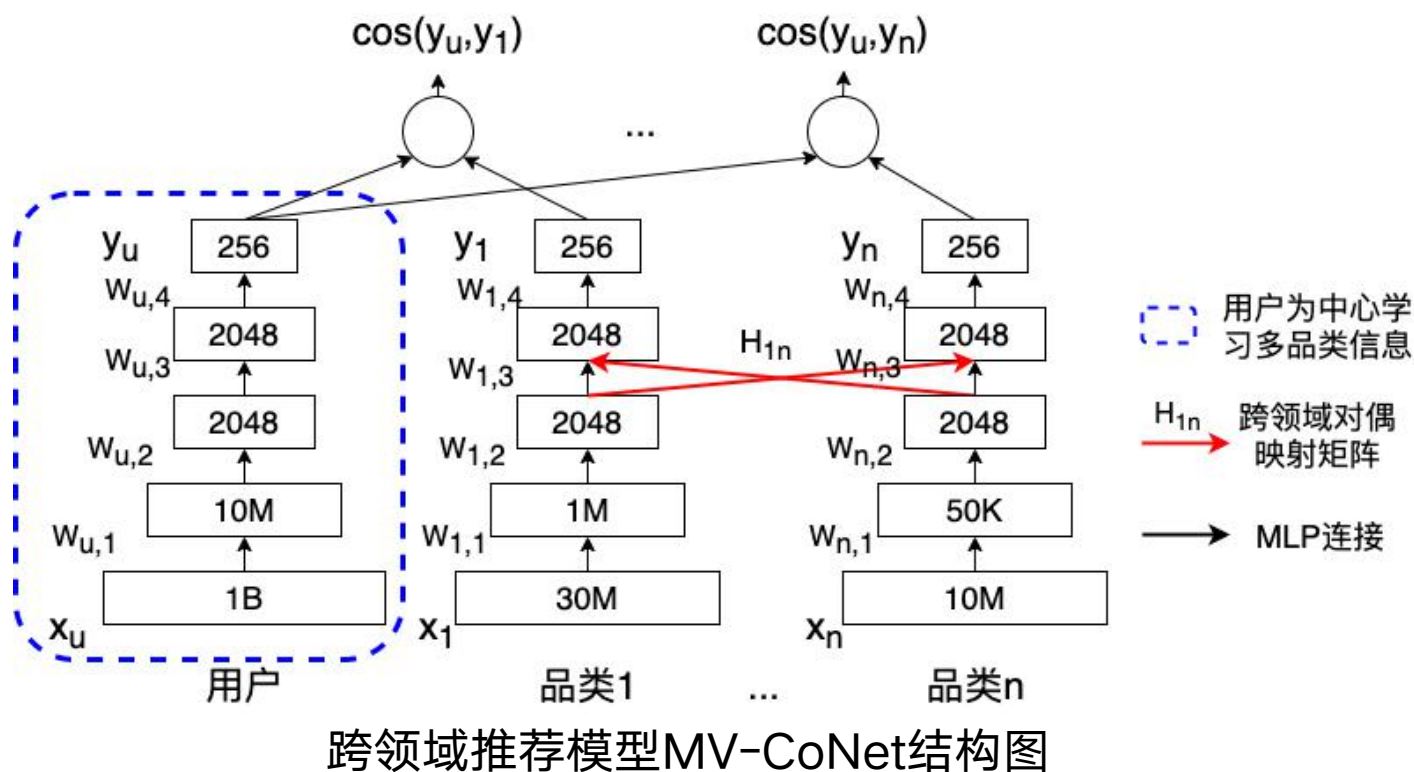
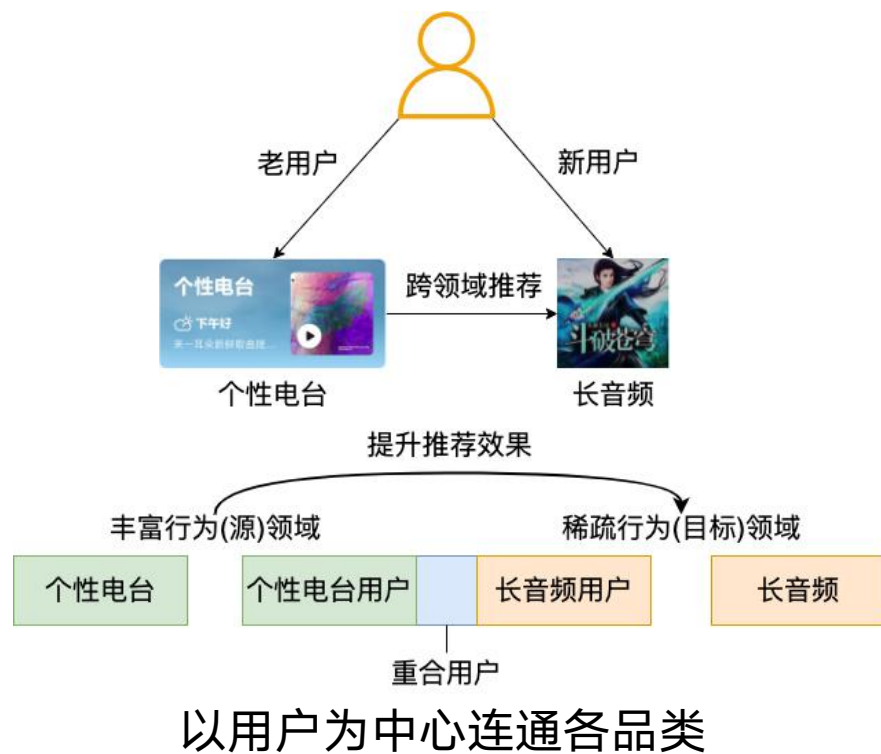


## 4.2 用户体验的多样性-跨领域推荐

问题：品类碎片化，不同品类缺乏协同

方案：跨领域推荐(cross-domain recommendation)

利用丰富行为的品类迁移学习到稀疏行为的品类，提升推荐效果。



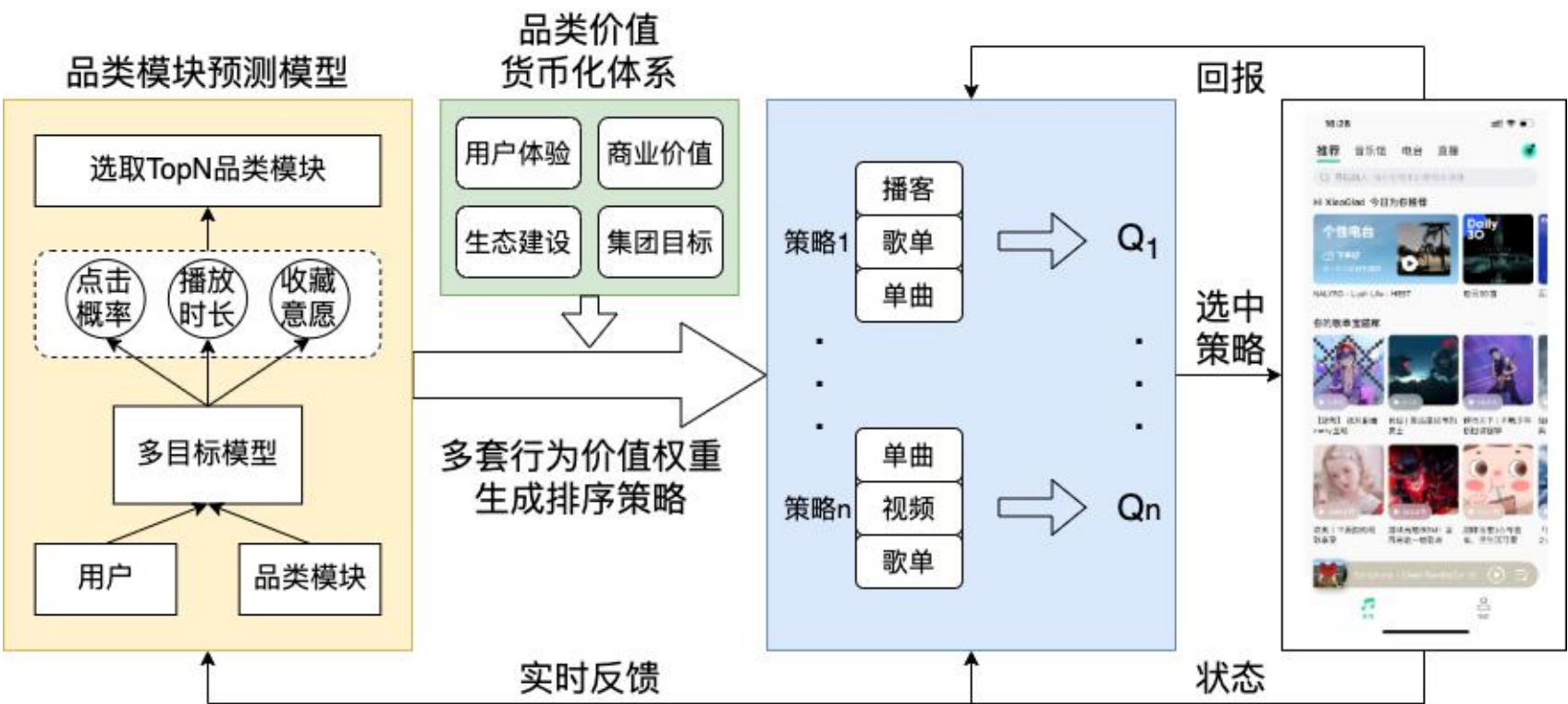


# 4.2 用户体验的多样性-模块个性化

方案：模块个性化，提升用户体验，获取更大流量收益。

$$E(\text{个性化策略}) = \sum_{p \in P} C_p * (1 + \alpha * E(p))$$

其中P为所有品类集合（歌曲、视频等）； C<sub>p</sub>为品类消费人数； E(p) 为品类价值； α为权重超参。



整体收益：

用户规模：首页DAU提升12%

用户体验：品类多样性提升20%

流量收益：整体品类价值提升17%







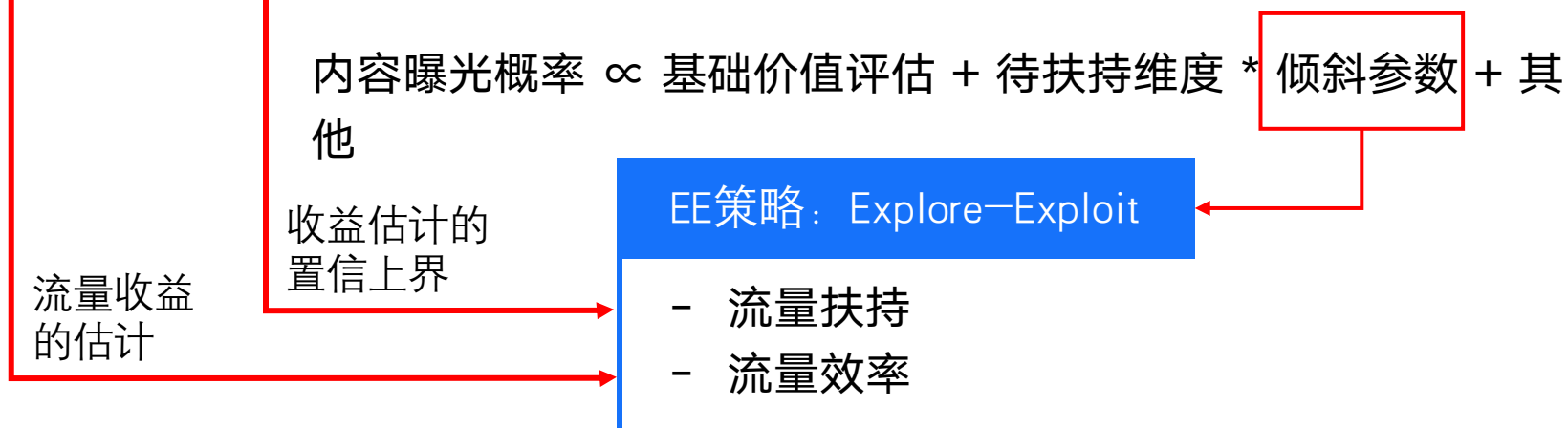
## 4.3 流量的利用和探索平衡问题

### ■ 案例：腾讯音乐人新歌or潜力歌曲扶持

- 采用Explore-Exploit探索模型，动态调整流量策略，兼顾“流量扶持”与“流量转化效率”的诉求

$$s = f(\alpha, z, x, y) = \begin{cases} z + \alpha * \sqrt{\frac{2\ln(x+1)}{y+1}} & y < N \\ z & y \geq N \end{cases}$$

其中： $\alpha$ 为scale因子， $z$ 是歌曲的实时价值评估分， $x$ 是大盘实时累计播放量， $y$ 是歌曲实时累计播放量， $N$ 是结束探索的播放量阈值





## 基础能力建设

基础指标体系  
——用户&内容&分发

内容理解与表征  
——离线实时评估&  
向量表征

用户画像与圈层  
——用户&圈层的  
内容喜好

## 冷启动优化

用户分层  
——活跃、人口属性、  
五维画像

用户冷启动  
——纵向联邦学习  
&DSSM/MMoE

内容冷启动  
——Graph&audio  
embedding

## 内容生态与多样性

马太效应  
——引入因果推断  
消除bias

用户体验多样  
——跨领域推荐和模  
块间排序

长尾音乐人动态扶持  
——流量效率的探索和  
利用平衡



创造音乐无限可能

CREATING ENDLESS  
OPPORTUNITIES WITH MUSIC