QQ音乐的个性化探索

腾讯音乐 / QQ音乐业务线 / 智能数据中心

Overview

01 | 关于音乐，关于用户

02 | 音乐个性化的思考和演进

03 | 广告个性化的尝试

04 | AI时代一些好玩的尝试

Overview

01 | 关于音乐，关于用户

02 | 音乐个性化的思考和演进

03 | 广告个性化的尝试

04 | AI时代一些好玩的尝试

QQ音乐 / 听我想听的歌

注册用户 **8亿**

DAU **1亿**

全民K歌 / 你其实很会唱歌

注册用户 **4.6亿**

主力军 90后用户

Play on all your devices

iOS/Android — PC — H5 — IOS/ANDROID 设备 — 车载 — 智能音响 — TV

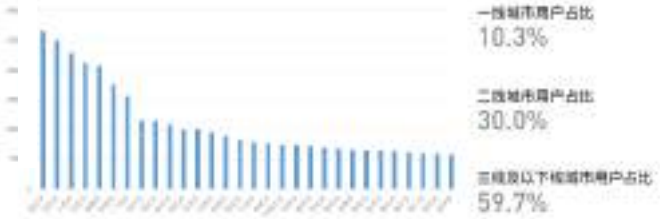
“音乐这件事儿“

我们的用户

腾讯音乐娱乐

一线城市阵地稳固，渠道下沉优势明显

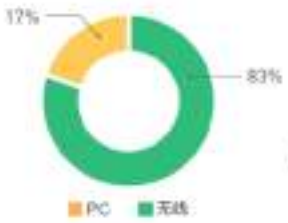
QQ音乐用户市场排行TOP30



超级数字音乐航母在此腾飞

智能终端 | 网页+客户端 | 无线+PC | 安卓+iOS | WIN+MAC 全平台布局

平台播放分布



PC | 无线占比

移动端播放分布

iOS | Android占比

我们的用户听什么？


Fantastic Baby
时间煮雨
我的歌声里
可惜没如果
滴答
泡沫
告白气球
依然爱你
泡沫
丑八怪
李白
IF YOU
我好想你
青花瓷
幻听
演员
小苹果
默
喜欢你
不再联系
平凡之路
一万个舍不得
七秒钟的记忆
闹够了没有
刚好遇见你
Faded
因为爱情
咱们结婚吧
我的好兄弟
Always Online

我们的用户听什么？

腾讯音乐娱乐

时间维度：Aug.2017 – Oct.2017

巅峰音乐


《刚好遇见你》 李玉刚


《演员》 薛之谦


《小苹果》 筷子兄弟


《李白》 李荣浩


《平凡之路》 朴树


《丑八怪》 薛之谦


《告白气球》 周杰伦


《凉凉》 杨宗纬|张碧晨


《Faded》 Alan Walker


《默》 那英

巅峰专辑


《意外》 薛之谦


《不良少年》 徐良


《绅士》 薛之谦


《模特》 李荣浩


《三生三世十里桃花》 原声


《我很忙》 周杰伦


《魔杰座》 周杰伦


《万有引力》 汪苏泷


《JJ陆》 林俊杰


《刚好遇见你》 李玉刚

巅峰艺人


周杰伦


薛之谦


陈奕迅


张杰


林俊杰


许嵩


G.E.M. 邓紫棋


BIGBANG


徐良


张学友

Overview

01 | 关于音乐，关于用户

02 | 音乐个性化的思考和演进

03 | 广告个性化的尝试

04 | AI时代一些好玩的尝试

推荐的那些事儿

长期
用户口碑，品牌调性

VS

短期
业务KPI，case by case


Café del Mar

个性化推荐，在路上

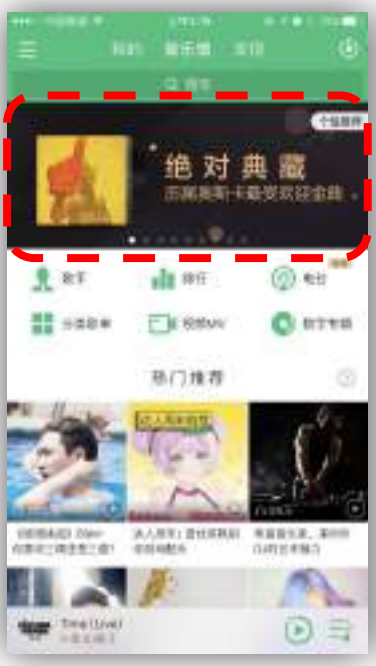
腾讯音乐娱乐

2011 – 2012

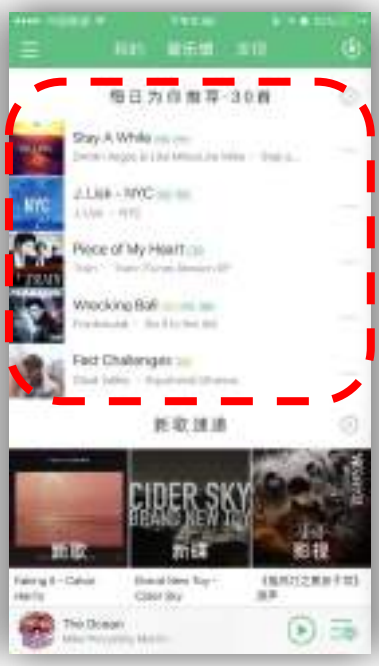
2013 – 2014

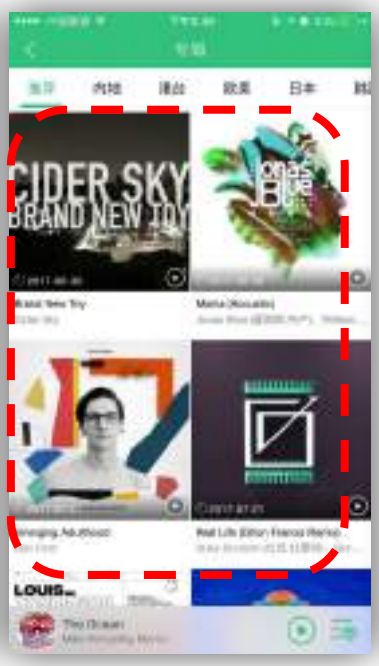
2014 – 2015

个性化推荐，依然在路上


绝对典藏
热门推荐

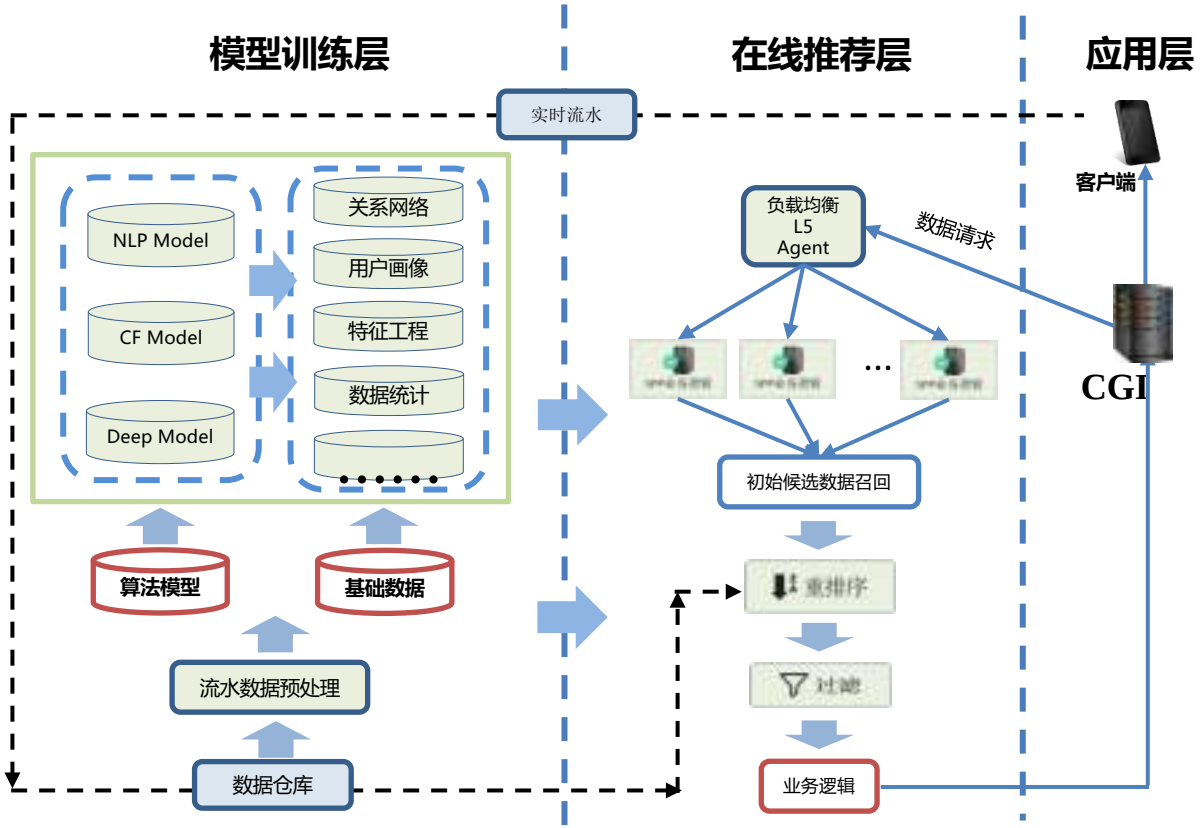

热门推荐


每日为你推荐·30首
Stay A While
Piece of My Heart
Wrecking Ball
Fact Challenges
新歌速递
新歌
新碟
电视


CIDER SKY
BRAND NEW TOY


北欧之声：来自极寒之地的歌声
女子摇滚乐队Sleater-Kinney送你一束干花
另一个角色的Childish Gambino

QQ音乐 – 个性化引擎 – 全景图



中心化推荐

- 规则引擎
- 热门音乐
- 新音乐

引入个性化推荐

- 基于内容/标签推荐
- CF
- 基于用户长期兴趣歌手
- 基于用户长期兴趣流派

优化个性化推荐

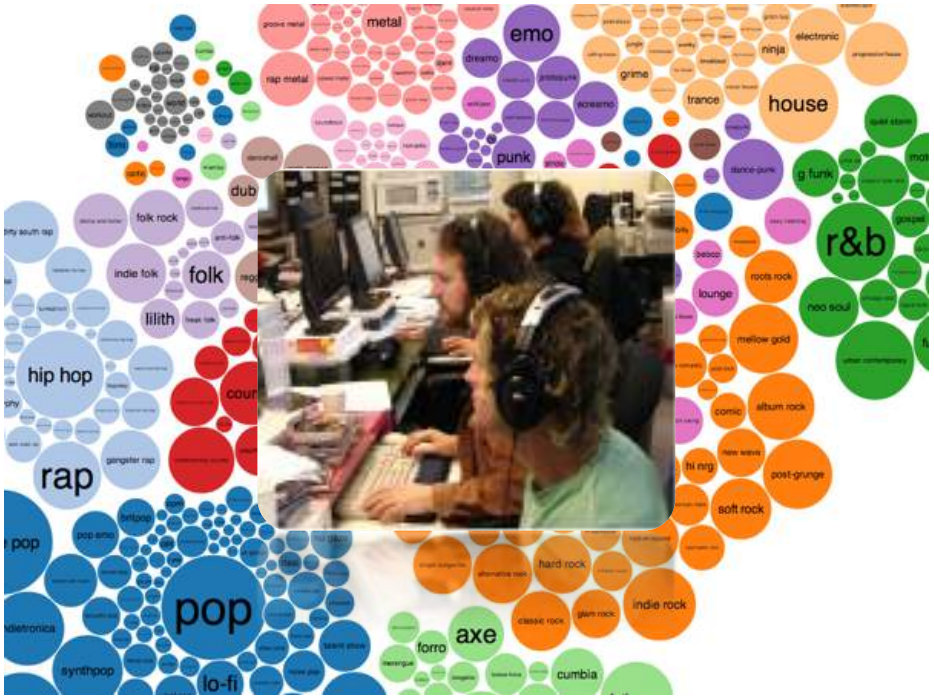
- 在线实时架构
- 冷启动优化
- 用户特征工程优化
- 多目标推荐优化
- 内容特征优化（歌单等文本类模型）

AI相关探索

- 音乐大数据挖掘
- 深度神经网络
- 图像理解

个性化引擎 – Part 0. 音乐理解

Pandora 专家人工标注



专业公司Gracenote以及学院派专业人员采用近2000种音乐元数据（流派、情感、主题标签）对每一首歌曲进行分析并标签化

个性化引擎 – Part 1. 歌单生态系统

腾讯音乐娱乐

歌单知识体系

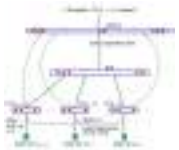
➢ 基础特征描述:

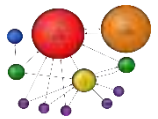
-语种,流派,歌手,年代分布...
-冷热程度
-播放流水,收藏流水

➢ 语义特征描述:

图片　文本　歌曲

PGC　UGC





歌单产品周期全面支撑

歌单投稿

日均机审占日审核量
70%+

广场排序

歌单收听数量提升
56%->87%

歌单推荐

音乐馆听歌显著增长

关联歌单

全面覆盖外部展示歌单

个性化引擎 – Part 2. 用户理解

多平台联动

- QQ，微信基础画像
- 腾讯视频
- 全民K歌
- 朋友圈/微博音乐分享数据等

海量用户行为数据挖掘

- DAU：1亿+
- 单用户日均操作数据：50亿+
- 每日歌曲播放：十亿级

用户特征

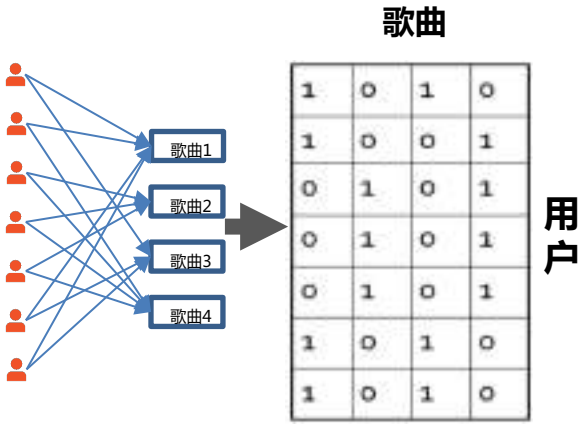
- 用户基础信息：性别、年龄、地域、学历…
- 音乐口味偏好：歌手、流派、语言、年代
- 音乐行为偏好：电台、收藏、下载、歌单、搜索
- 平台行为：新增、留存、回流、活跃…

用户特征挖掘算法

- 听歌及操作流水，时间衰减模型
- 自然语言处理：对文本数据，如评论、歌单的标题挖掘
- 噪声过滤：SPAM等
- 监督学习：利用LR、GDBT等模型进行用户喜好预测

个性化引擎 – Part 3. CF Model

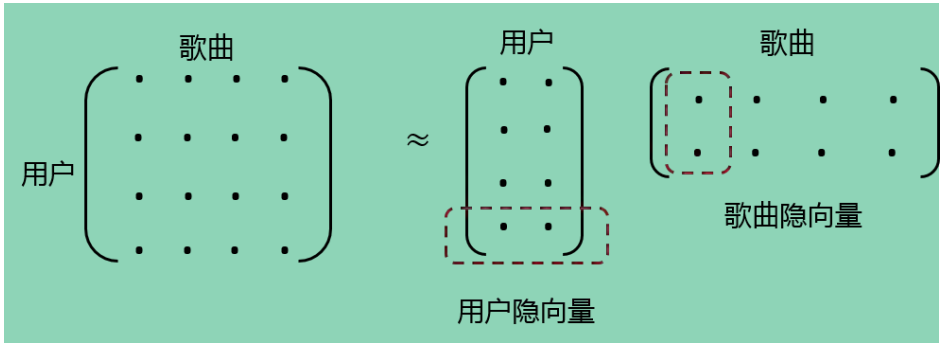
浅层协同：
利用用户的听歌行为数据构建user-item矩阵，求取相似歌曲或相似用户，无须领域知识。



$$\mathbf{sim(i,j)} = \frac{\mathbf{A \cap B}}{\mathbf{A \cup B}}$$

- 模型简单，准确度高，且可解析性好

隐因子模型:
Latent Factor Model：利用矩阵分解方法，求出用户和歌曲的隐特征向量。



目标函数: 优化均方误差RMSE，使得预测评分与真实评分的均方误差最小

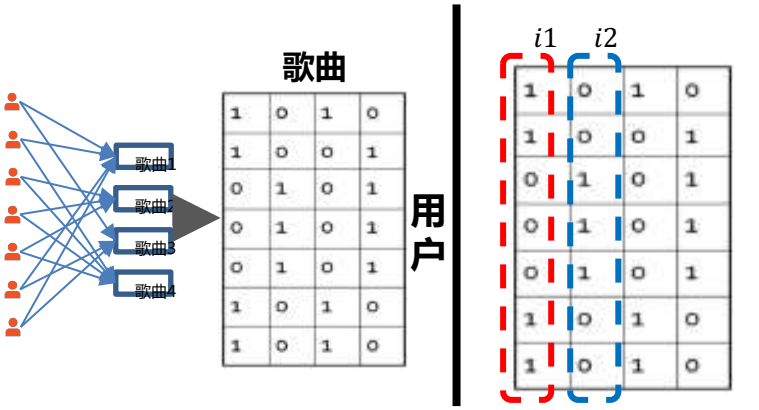
$$\min_{x^*,y^*} \sum_{u,i} c_{u,i}(p_{ui} - x_u^T y_i - \beta_u - \beta_i)^2 + \lambda\left(\sum_u \beta_u^2 + \sum_u \beta_i^2\right)$$

p_{ui}	一个二值变量，当$r_{u,i} > 0$时，直接置为1，否则置为0；其中$r_{u,i}$是用户u对歌曲i的得分
β_u, β_i	分别代表用户与歌曲的隐特征向量的偏移量
x_u, y_i	x_u代表用户的隐特征向量，y_i代表歌曲的隐特征向量
$c_{u,i}$	置信度，由于隐因子模型考虑的是隐式反馈，因此需要一个置信度来表示得分置信度 $c_{u,i} = 1 + \alpha * r_{u,i}$

个性化引擎 – Part 3. CF Model的挑战

●协同模型(CF Model)的挑战

- 仅能发现浅层特征。
  - 缺点一：推荐的歌曲风格单一，缺乏新鲜感
  - 缺点二：只挖掘浅层的特征，无深层的特征



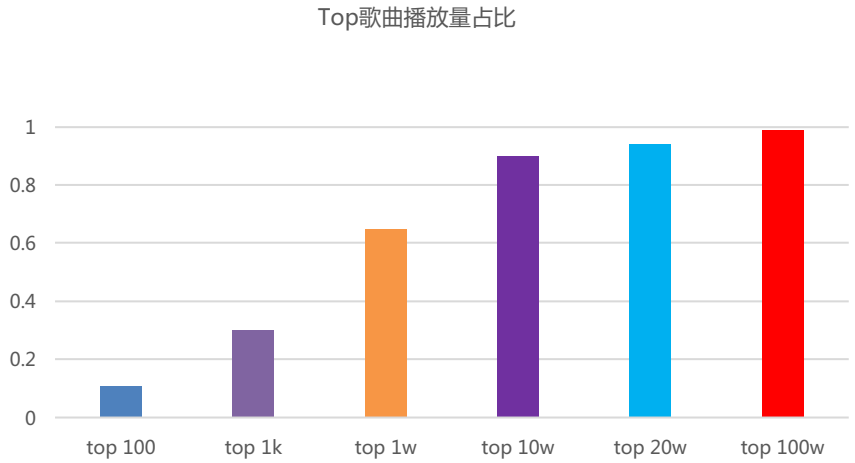
$$sim(i,j) = \frac{A \cap B}{A \cup B}$$

$$sim(i1, i2) = 0\ ?$$

$$sim(i1, i4) = \frac{1}{7}$$

$$sim(i2, i4) = \frac{3}{4}$$

- 马太效应明显，Top100万歌曲占据了总收听量的90%+；
- 基于用户行为召回的数据，多以热门数据为主，如何跳出热歌圈子，挖掘长尾歌曲？
- 亿级用户的协同计算性能问题。



个性化引擎 – Part 3. CF Model的挑战

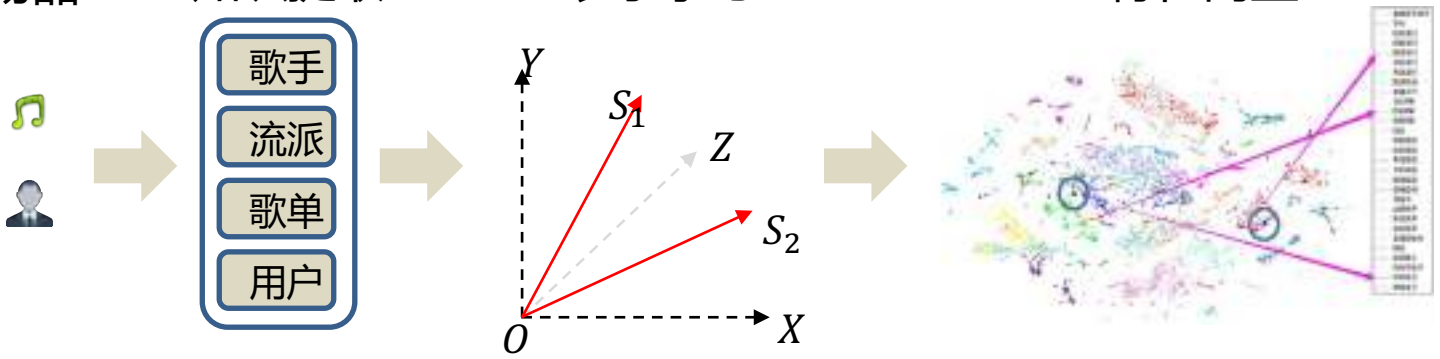
- **相似度计算公式小优化**
  - 引入IUF(Inverse User Frequency)

By John S.Brees, David Heckerman， Carl Kadie <Empirical Analysis of Predictive Algorithm for Collaborative Filtering

$$sim(i,j)=\frac{\sum_{u\in N(i)\cap N(j)}\frac{1}{\log(1+|N(u)|}}{\sqrt{|N(i)||N(j)|}}$$

- **离线计算度量优化**

物品　　知识提取　　表示学习　　特征向量

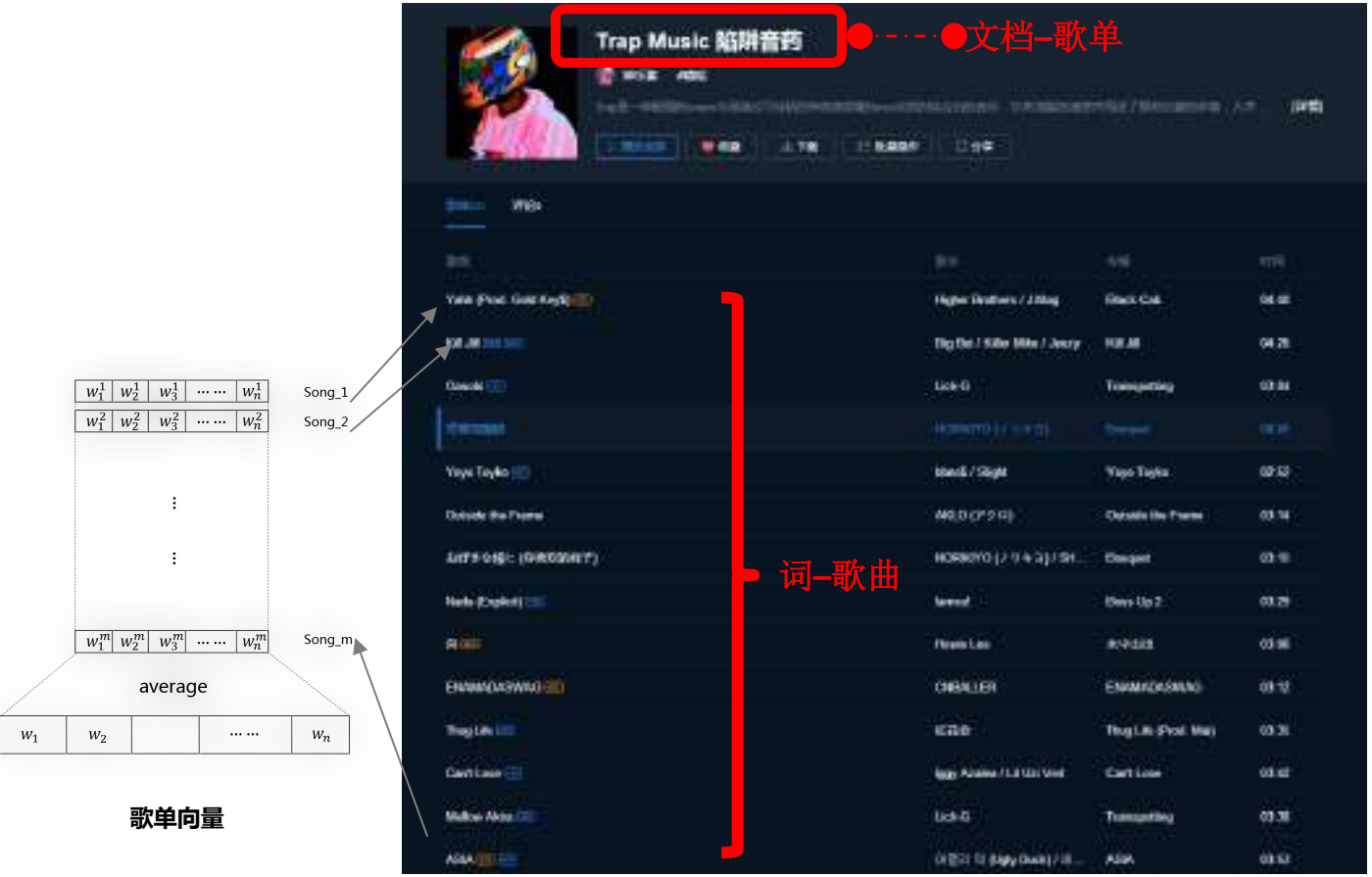


如何评价不同的特征向量表示的质量：

- 线上环境评测：相似单曲电台等实测场景。人均听歌时长，单曲听歌时长等。
- 模型的评测：特征向量的表示，应该能够使得在同一个流派下，歌曲之间的特征向量距离尽量接近。

$$min\left(\frac{1}{m}\sum_{v}\|s_i-s_j\|^2\right)$$

个性化引擎 – Part 4. NLP Model



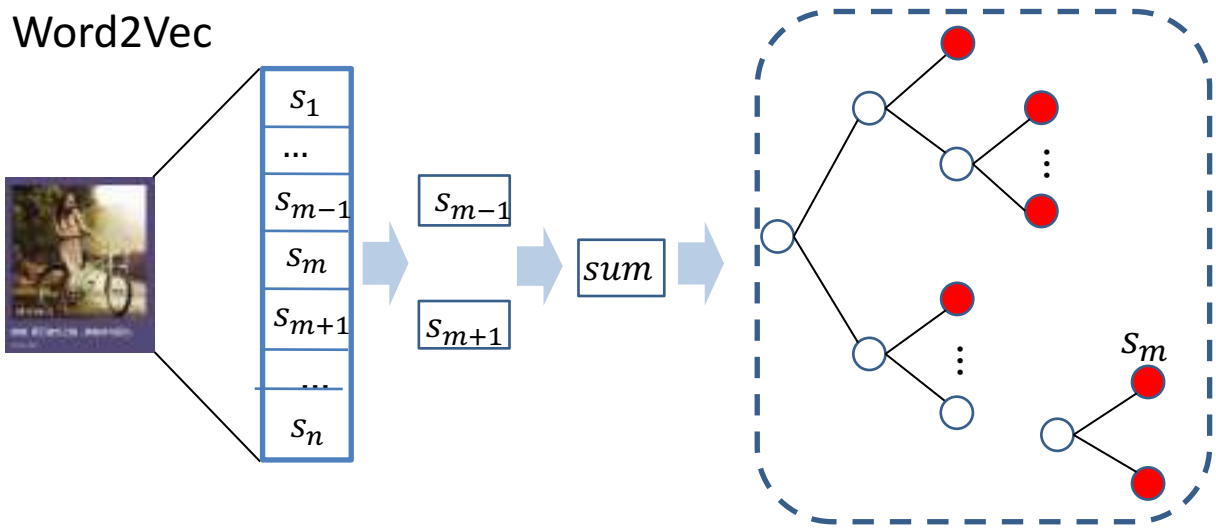
与CF Model相比，NLP Model的应用场景多种多样，既可以用于作为推荐数据召回建模 ，也可以用于特征提取建模

主动热度降权：
歌单数据中存在大量的小众歌曲，有利于进行长尾推荐；并且受到大盘听歌流水的影响较小，降低了噪音对模型的训练的影响

NLP Model的主体思想：
将歌单作为文档，通过word2vec求取每一首歌曲单词的词向量表示

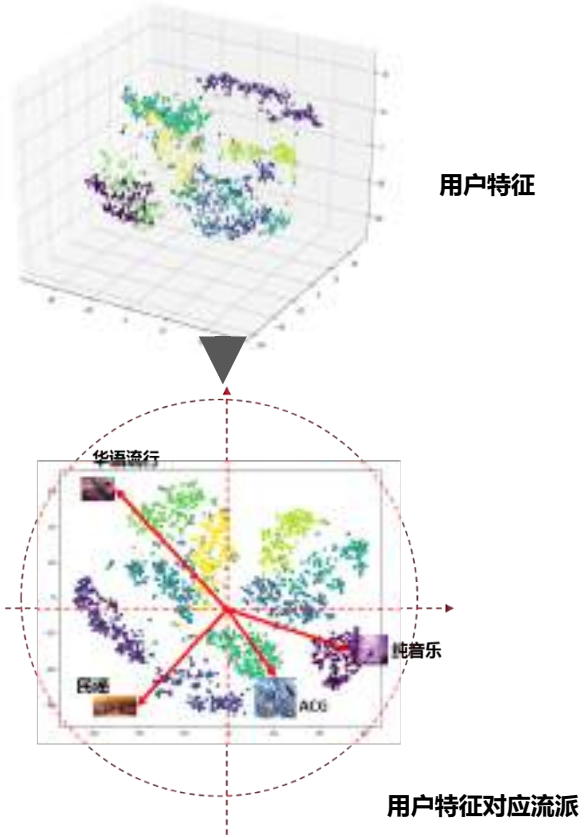
数据增强：
精选歌单质量好，
但数量少，通过组合歌单来扩展语料库

个性化引擎 – Part 4. NLP Model



Manifold Learning

利用t-SNE对高维的用户特征和歌曲特征进行降维:



- 将歌单作为文档，通过word2vec求取每一首歌曲的词向量表示，倾向于选择热门的歌曲做负样本进行训练 (negative sampling)。越热门的歌曲，离根节点越近。

  用户没有点击某一首歌曲，通常有两种原因：
  一是不知道有这首歌曲;
  二是不喜欢这首歌曲
  对于热门歌曲来说，显然第二种的可能性更高，这也是将热门物品作为负样本的合理性原因

个性化引擎 – Part 1~4的一些阶段效果

腾讯音乐娱乐

- **一些阶段效果**

➢ **产品指标的提升**

- 听歌人数：Android，Iphone平台提升20%+
- 人均听歌时长： Android，Iphone平台提升15%+

➢ **一些用户好评**

个性化引擎 – Part 5. DNN Model

Deep Neural Networks for YouTube Recommendations

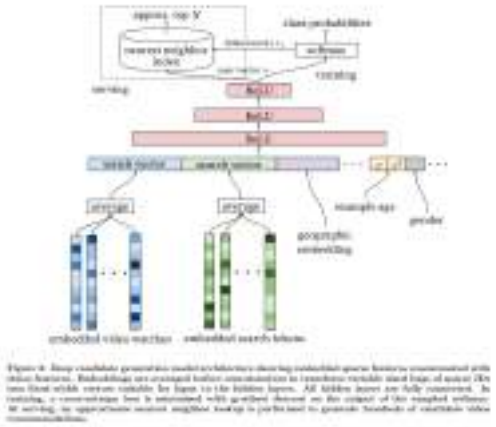
Paul Covington, Jay Adams, Emre Sargin
Google
Mountain View, CA
{pcovington, jka, msargin}@google.com

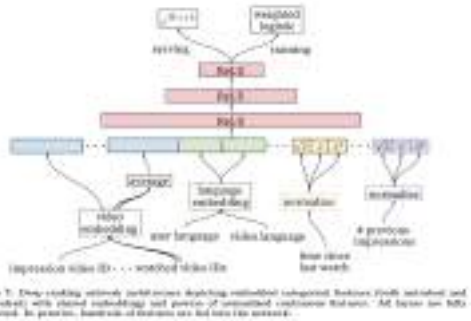
ABSTRACT

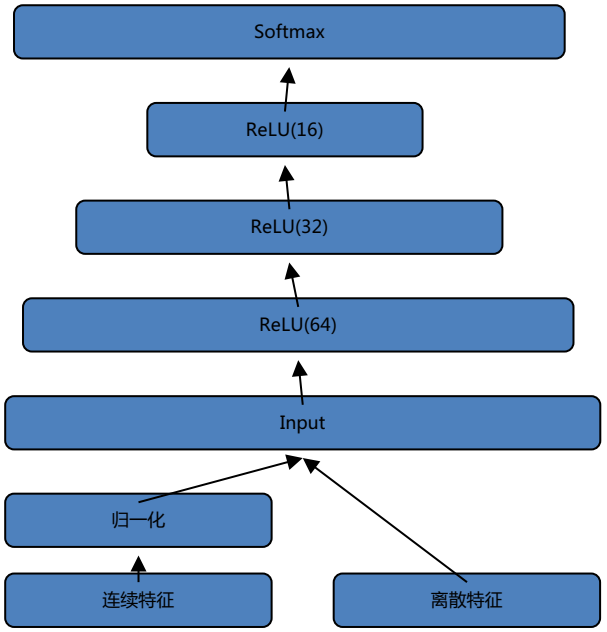
Keywords

1. INTRODUCTION

Figure 1: Recommendations displayed on YouTube mobile app home.







Overview

01 | 关于音乐，关于用户

02 | 音乐个性化的思考和演进

03 | 广告个性化的尝试

04 | AI时代一些好玩的尝试

QQ音乐 – 在线广告 *“广告，也可以是生活的一部分”*

MusicBoss精准营销平台

- 产品运营自助配置广告；
- 效果追踪等一站式闭环管理；
- 定向投放与模型投放相结合；

用户包定向筛选	广告投放	活动模型
5000	1000	100

口冷启动 ：使用CF模型，针对特征标签少的用户及活动进行冷启动；

口特征拓展：利用word2vec等算法，进行特征维度拓展，并计算人群lookalike；

口活动推荐：排序模块使用Xgboost等模型，进行特征离散化及线上实时预测；

口线上优化：使用FTRL等算法，根据用户的反馈数据实时优化模型参数；

QQ音乐 – 个性化在线广告 *“广告，也可以是生活的一部分”*

我们的在线广告探索：

生长阶段

2015年前

QQ音乐率先推出会员制，数字专辑等多种付费模式，推动音乐行业正版化，内部业务广告需求增长；同时逐步开放外部合作广告。**缺乏统筹，野蛮生长。**

产品化阶段

2015-2016年

逐步开始规范各个广告位的投放内容和形式。
纯人工运营，后台逐个需求开发。

平台化阶段

2016-2017年

搭建广告统一管理平台

搭建了音乐广告管理平台，对所有广告位进行统筹管理，整合广告业务相关功能，引入推荐算法。
平台化整合各个功能，完成了推荐算法、数据分析、投放策略上的迭代升级。

自动化阶段

2017年

加入了更多自动化功能，包括新增广告自动建模，流量自动分配等。
产品只需配置新广告，制定投放基本策略，平台在投放周期内会进行流量控制。

QQ音乐 – 个性化在线广告 *“广告，也可以是生活的一部分”*

腾讯音乐娱乐

平台化阶段 2016-2017年

APP内部广告

数字专辑

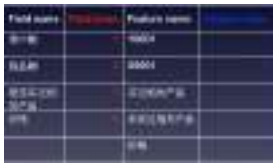
付费音乐包

外部合作

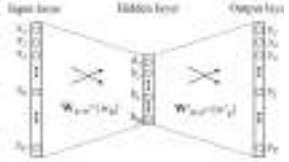
会员推广

计算模块—核心算法变迁

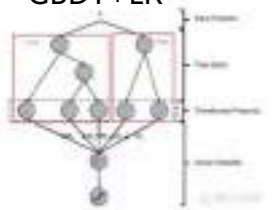
FM因子分解机



item2vec



GBDT+LR



XGBOOST

排期模块—运营与模型结合

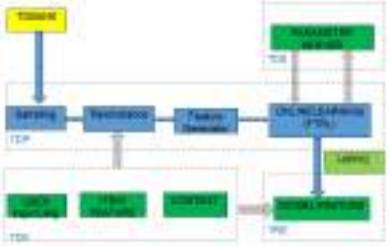
- 定向投放
- 模型投放
- 定向+模型
- 频控投放

分析模块—全流程实时化

画像实时分析—Hermes（腾讯）

用户画像近10亿数据(约10TB)分析对比		
	hadoop离线分析查询	实时多维分析
Count	15min+	1~3s
Sum	25min+	5~6s
Group by	26min+	1~3s

模型参数实时调整—FTRL



投放效果实时监控—TRC

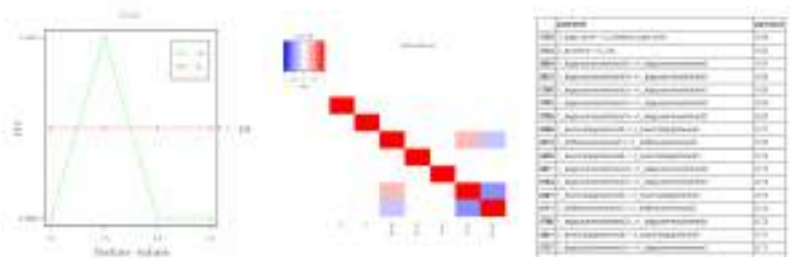
QQ音乐 – 在线广告 *“广告，也可以是生活的一部分”*

广告模型自动建模和优化：

新上线广告，进行短期随机投放后，可**自动建模；**
已上线广告，每隔一段时间会根据反馈数据，自动调整模型；
自动化任务流如下：

Assemble
特征集成，负责将样本和新特征进行集成
Transform
特征转换，负责做特征常用转换，比如特征离散，特征交叉，tf-idf
Criteria
特征评估指标计算，包括entropy-ig, giniindex, entropy-igr, symmetry-uncertainty等
Model
模型评估指标计算，包括auc、logloss、rmse等，以及输出特征全局重要度、树模型等

自动输出变量探索报告



流量分发自动控制

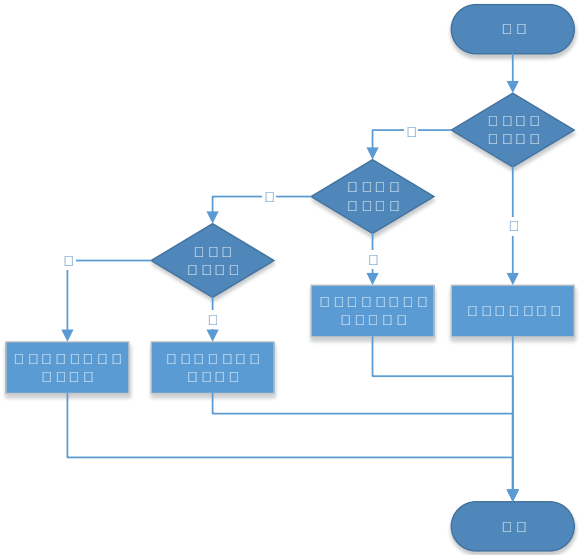
为了**处理人工运营、定向投放、智能推荐**几种不同投放方式的矛盾，平台可进行不同投放方式的组合，例如：

1）定向人群投放单一指定广告。

2）定向人群进行多个广告的智能排序推荐；非定向人群按优先级投放。

3）定向人群投放单一指定广告；非定向人群多个广告智能排序推荐。

当流量进入多个广告智能推荐分支时，又会根据模型效果，进行动态的流量调整。



Overview

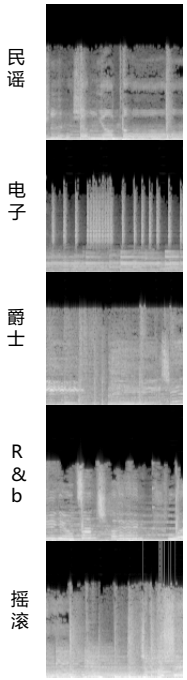
01 | 关于音乐，关于用户

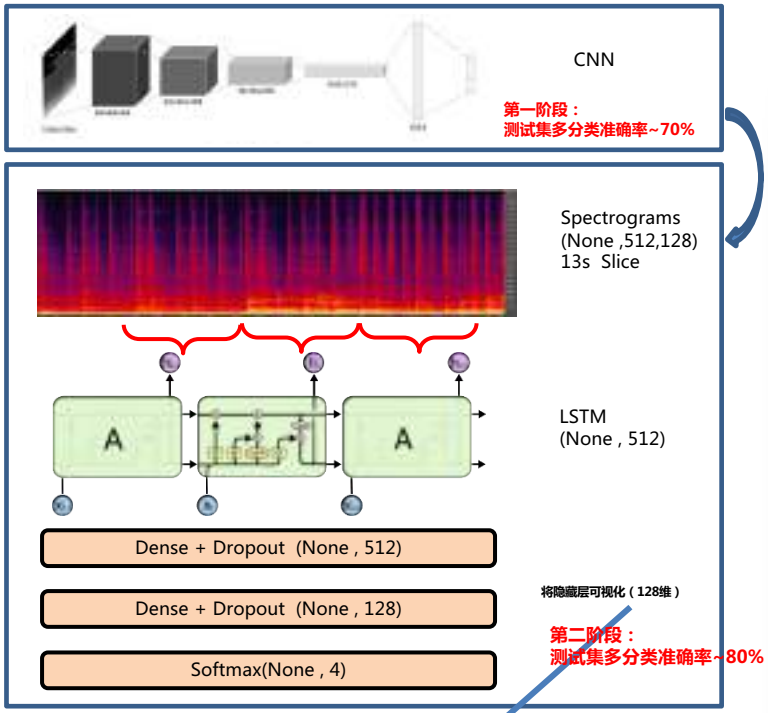
02 | 音乐个性化的思考和演进

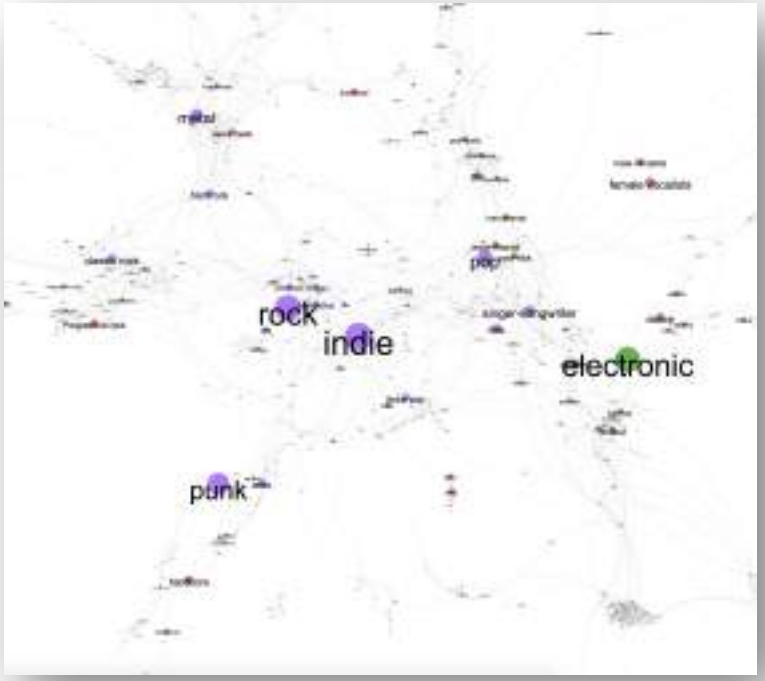
03 | 广告个性化的尝试

04 | AI时代一些好玩的尝试

AI时代，一些好玩的尝试







By Echonetst

AI时代，一些好玩的尝试

增量关联

距离　聚类　GAN

统一的特征空间

主题：古典、乐器　　实体：钢琴

抽象特征：[0.726，0.032，0.438，.....]　　情绪：中性

CNN　NLP　图谱

存量分析（举例）

封面：

歌单：《纯音｜50首轻缓闲适钢琴曲》

歌曲：《卡农》

相关实体：帕赫贝尔

THANKS

QQ MUSIC
DATA LAB