



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





# 全民K歌 推荐系统

基于社交&UGC内容的个性化推荐

腾讯音乐集团高级算法工程师 黄昕



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



# Contents 大纲

---

1

## 整体介绍

业务介绍 系统架构 算法架构

2

## 特征工程

社交特征 ID特征 内容特征

3

## 推荐排序

协同算法比较 排序模型比较

4

## 平台建设

整体流程 特征处理



十年架构 成长之路





# 整体介绍

业务介绍

系统架构

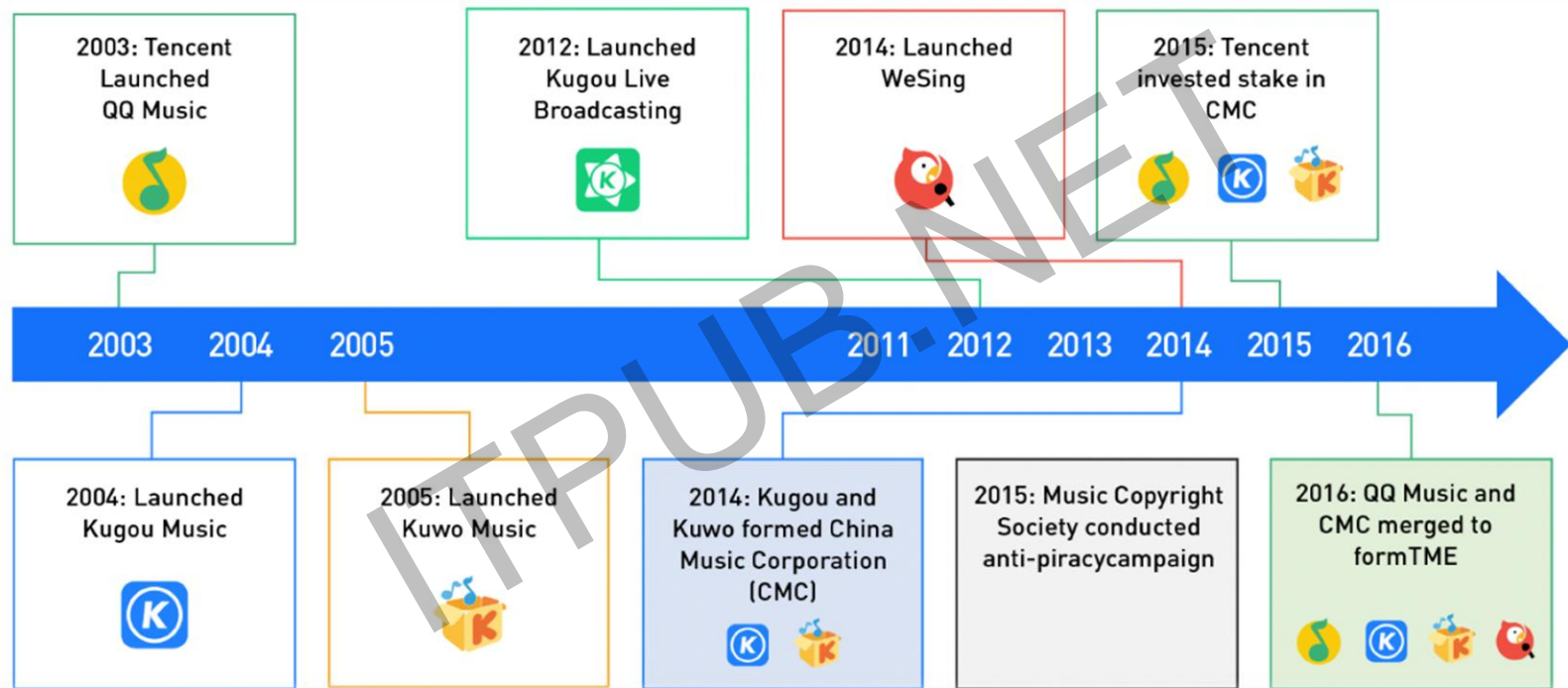


SACC

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



## 腾讯音乐集团发展历史





全民K歌



K歌



直播



社交



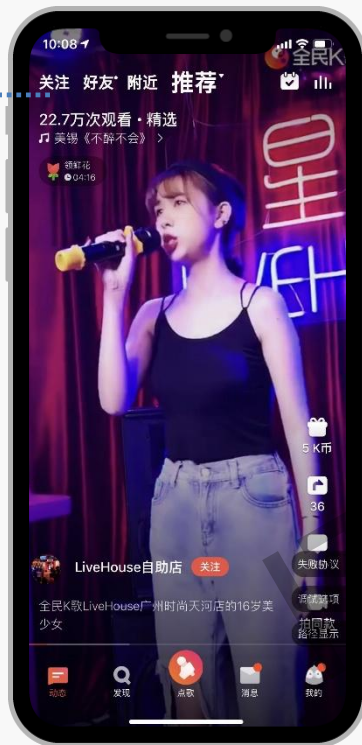
2017年腾讯名品堂



## POPUP

基于兴趣

基于用户音乐偏好+内容偏好做个性化**视频**推荐

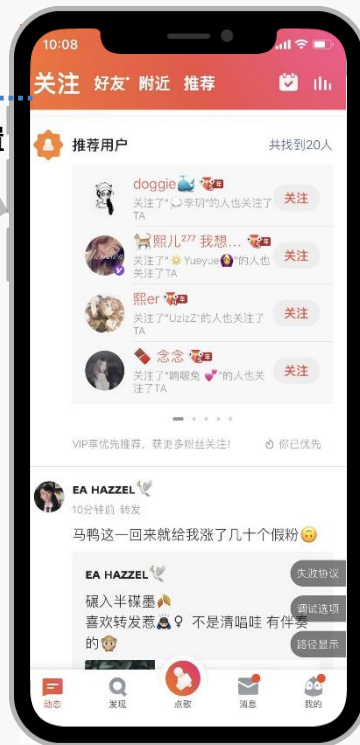


## FEED

基于关注

基于位置

多类型推荐：  
基于用户发布/转发内容  
包含推荐**音频/视频/好友**







# 特征构造

社交特征


ID特征


内容特征





## FaceBook EdgeRank

$$\Sigma = U_e \times W_e \times D_e \quad ?$$

 Rank

 亲密度 ?

 权重

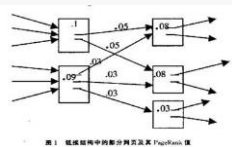
 时间衰减

## 1.1、pagerank影响力：

使用用户间关注+听歌+送礼的有向边构造图，计算每个节点的影响力：

使用以上方法进行影响力评估的好处：

- 1) 排除了僵尸粉丝、不互动粉丝的影响。
- 2) 考虑了粉丝的质量，例如一些用于刷量的小号，不会传递出较高的影响力。



## 1.2、personal pagerank影响力：

从用户u对应的节点 $V_u$ 开始游走，游走到任意节点按照概率 $\alpha$ 决定继续游走， $1-\alpha$ 的概率从 $V_u$ 重新游走。经过多次游走后，得到其他用户对用户u的个性化影响力。(注意：初始化时只有 $V_u$ 为1，其余点为0)

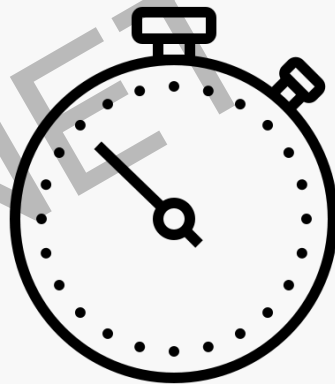
$$\begin{aligned} ppr(i) &= (1 - \alpha)r_i \\ &+ \alpha \sum_{j \in in(i)} \frac{ppr(j)}{|out(i)|} \end{aligned}$$

$$r_i = \begin{cases} 1 & i = u \\ 0 & i \neq u \end{cases}$$

- 1) 全局影响力高，但和我社交/兴趣圈较远的用户影响力降低。
- 2) 和我点对点存在边，但和我社交/兴趣圈其他人不存在变的影响力降低。

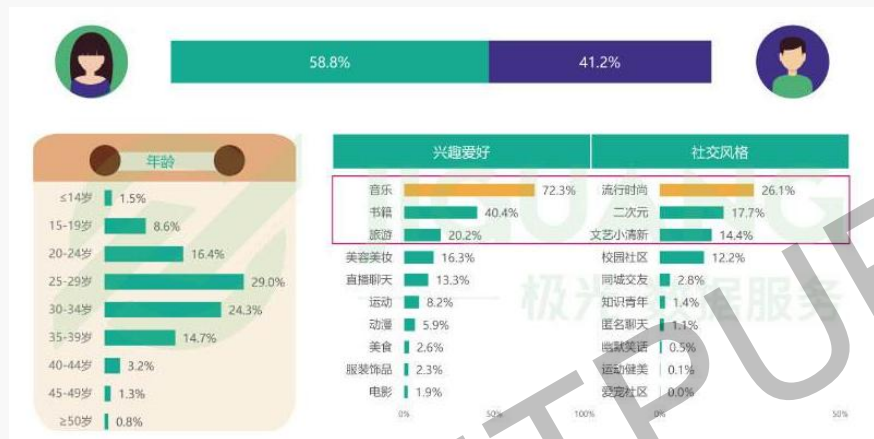


人力成本



实时性

## 传统的画像计算



10代表男 01代表女  
100000代表视频A 010000代表视频B

10 X 100000 点击

01 X 010000 点击

10 X 100000 点击

10 X 010000 不点击

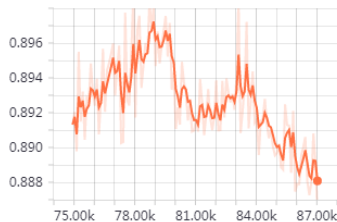
01 X 100000 不点击

01 X 010000 点击

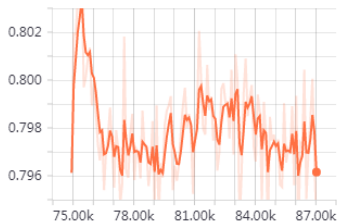


dnn

dnn/dnn/hiddenlayer\_0/fraction\_of\_zero\_values



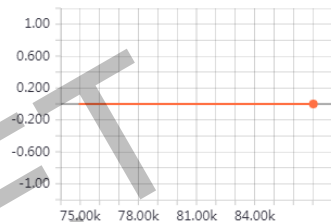
dnn/dnn/hiddenlayer\_1/fraction\_of\_zero\_values



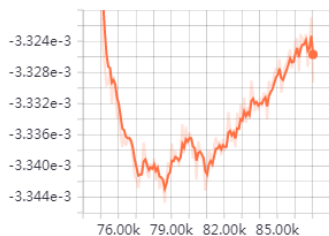
dnn/dnn/hiddenlayer\_2/fraction\_of\_zero\_values



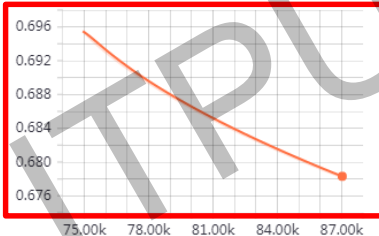
dnn/dnn/logits/fraction\_of\_zero\_values



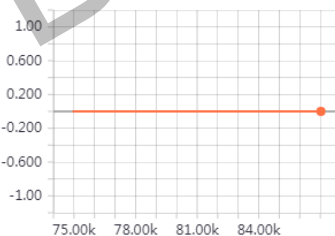
linear/bias



linear/fraction\_of\_zero\_weights



linear/linear/fraction\_of\_zero\_values

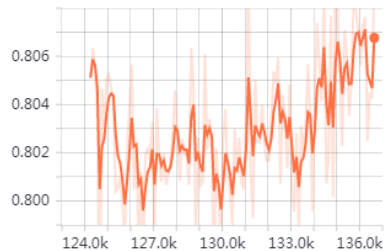


dnn

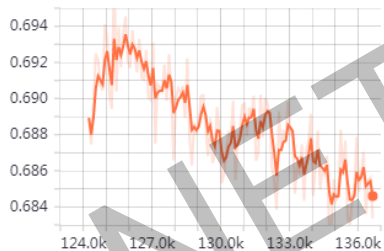
dnn/dnn/hiddenlayer\_0/fraction\_of\_zero\_values



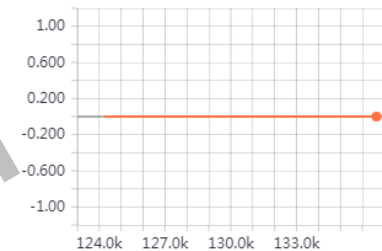
dnn/dnn/hiddenlayer\_1/fraction\_of\_zero\_values



dnn/dnn/hiddenlayer\_2/fraction\_of\_zero\_values



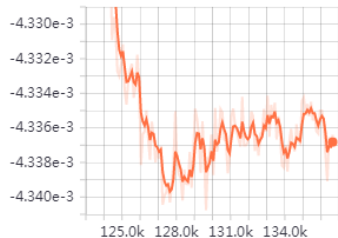
dnn/dnn/logits/fraction\_of\_zero\_values



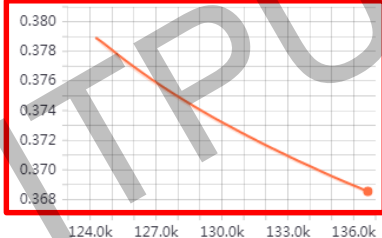
linear

36%系数为0，再低可能出现严重的hash冲撞

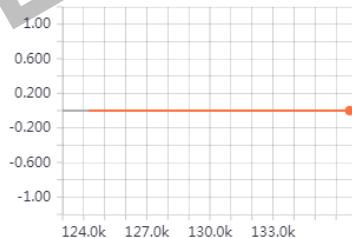
linear/bias

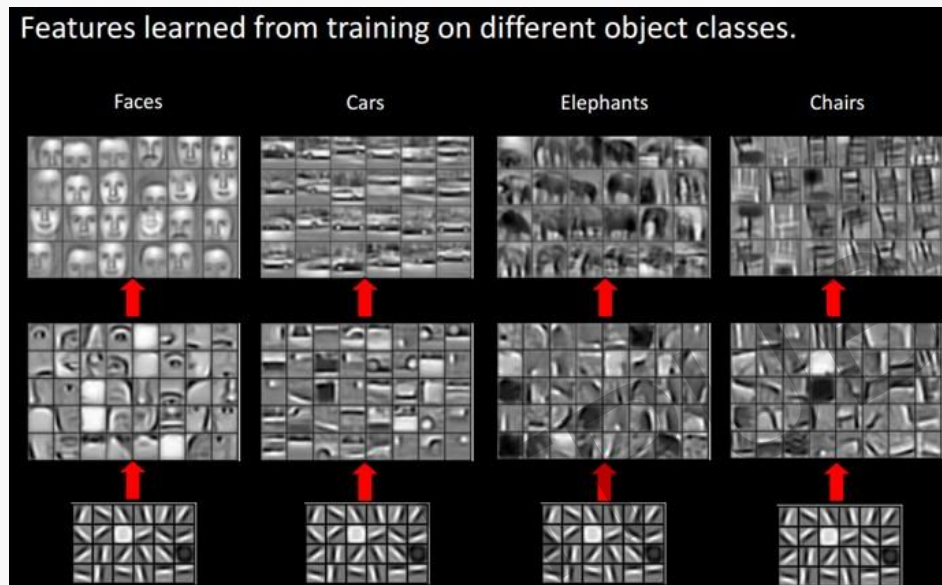


linear/fraction\_of\_zero\_weights



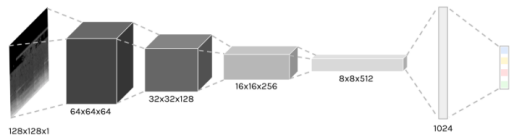
linear/linear/fraction\_of\_zero\_values





深度学习计算推动  
隐特征的发展



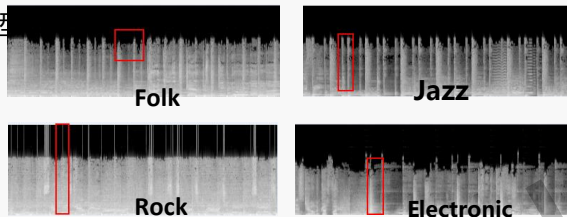


CNN

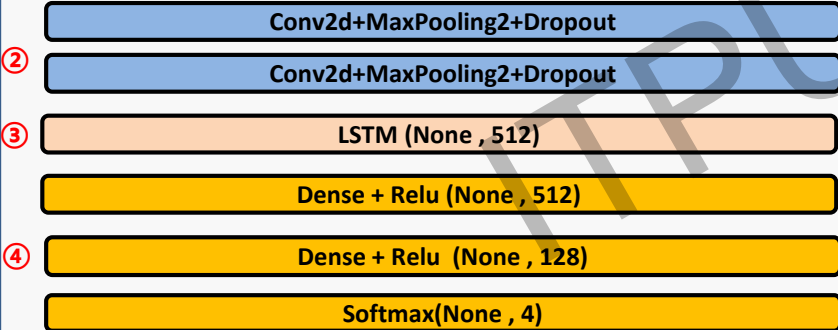
测试集多分类准确率67%

- 一般图片分类具有invariance(不变性)；音频图有时域和频域的物理意义。
- CNN通过filter size获取前后信息，但是受限于size大小，long dependence方面不如LSTM。

方案一：使用时序模型



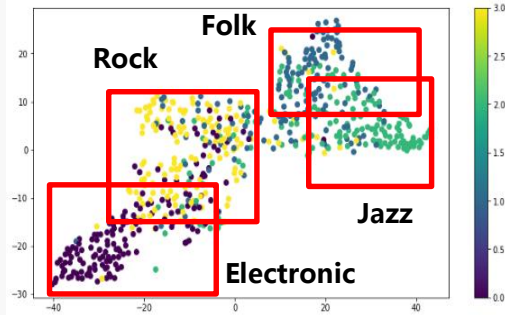
Spectrograms  
(None, 512, 128)  
10s Slice  
20ms per pixel



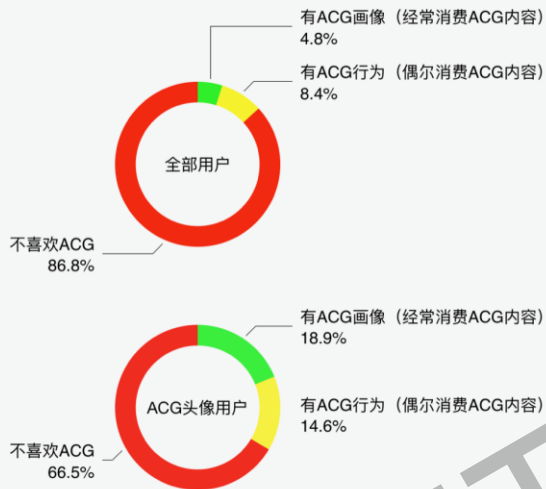
- 1) 卷积层构造局部特征
- 2) lstm构造序列特征
- 3) 将隐藏层可视化 (128维)

测试集多分类准确率73%

TSNE抽样可视化结果



## 用户头像



## 视频关键帧





# 推荐排序

协同算法比较

推荐算法比较

## item-CF ( 2014-2015 )

$$w_{ij} = \frac{|N(i) \cap N(j)|}{|N(i)||N(j)|}$$

$$w_{ij} = \frac{\sum_{u \in N(i) \cap N(j)} \frac{1}{\log(1 + |N(u)|)}}{|N(i)|^\alpha |N(j)|^{1-\alpha}}$$

物品被点赞多的降权

## SVD ( 2016 )

$$r_{ui} = \mu + b_i + b_u + q_i^T p_u$$

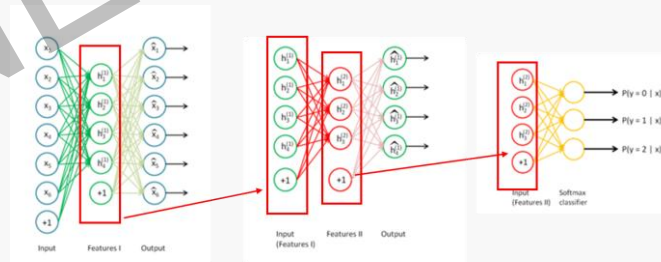


LFM

$$r_{ui} = q_i^T p_u + x_u^T y_i$$

利用用户画像和物品标签进行优化

## 稀疏编码 ( 2016-2017初 )



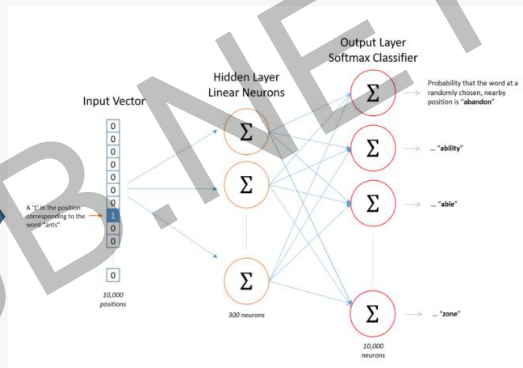
$$c_{ui} = 1 + \alpha \log(1 + \epsilon r_{ui})$$



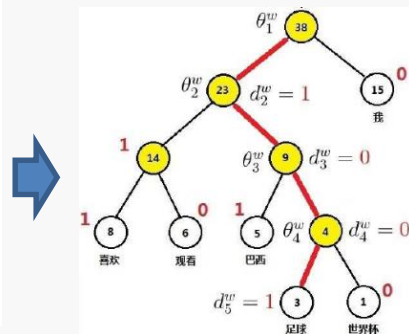
**Input :**  
骑单车去海边找你

(窗口设2)

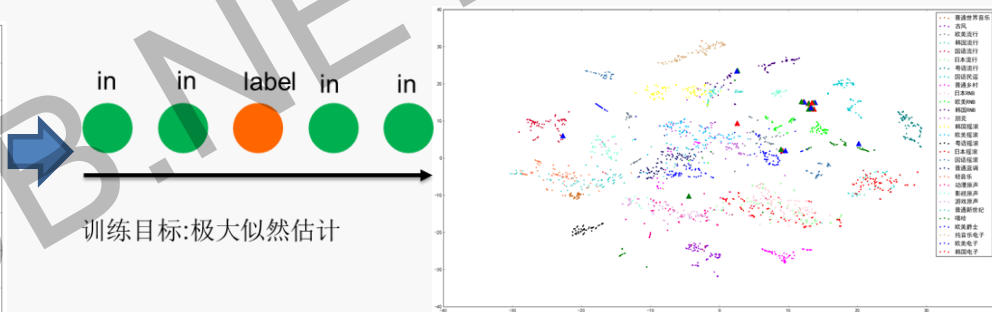
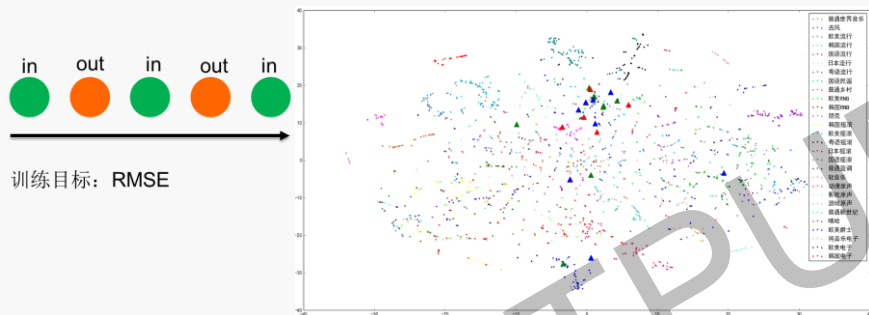
**Output :**  
海边,  
海边的卡夫卡,  
海边细浪,  
遗忘的海边



hierarchical softmax



# SVD VS WORD2VEC



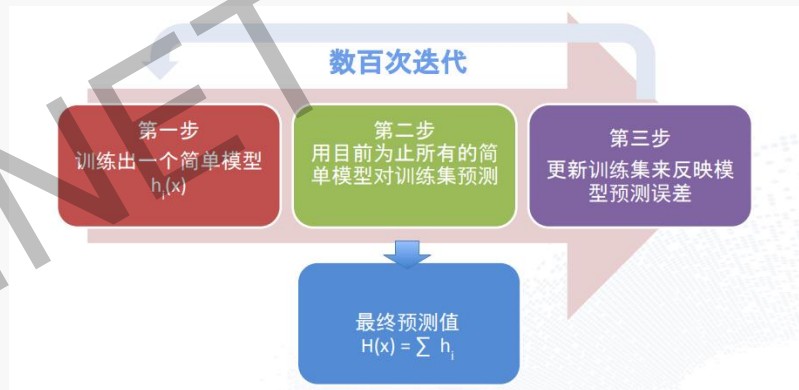
# 树模型 -> 深度模型

效果稳定，适合入门      减少人工，效果上限高



## XGBoost (树模型) 特点：

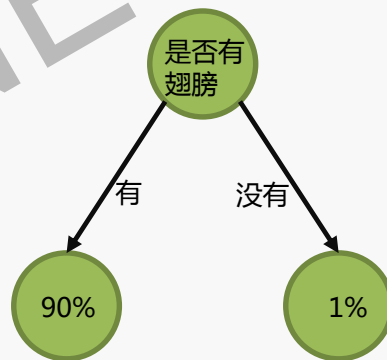
- 1) 基础的原理：多个弱分类器逐步把分类效果提升；
- 2) 需要人工构造大量特征；
- 3) 效果稳定，训练和预测的效率都比较高，适合作为推荐系统第一个模型版本；



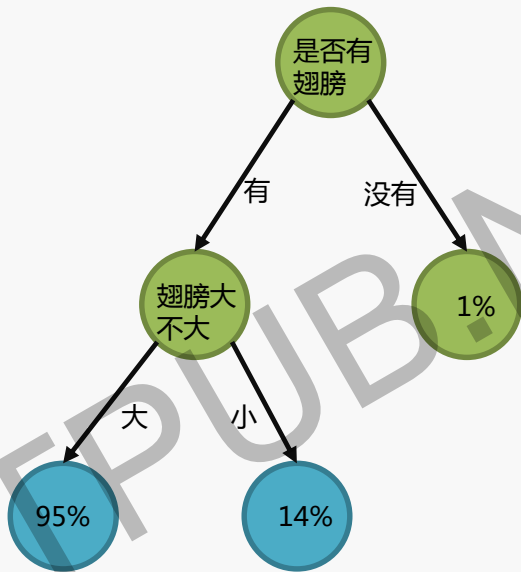


如何分类动物是否会飞？

有**翅膀**的动物会飞。



注：案例来自于tensorflow2017峰会



- 1) 需要不断找到**新的弱分类器**减少 badcase;
- 2) 需要有**标签** ( 是否有翅膀 , 大/小 ) ;

## 记忆+归纳

记忆 ( Wide ) : “老鹰会飞” “鹦鹉会飞”

归纳 ( Deep ) : “有翅膀形状的动物会飞”

记忆+归纳 ( Wide&Deep ) : “有翅膀形状的动物会飞, 但是企鹅不会飞”

1) deep层帮助我们从隐语义归纳特征, 减少标签的依赖;

2) wide部分记忆部分case, 减少对弱分类器的依赖。



# ④ 平台建设

整体流程

特征处理

早期数据处理格式：

```
20180917,2115614647,2118919384,2118919384_1537155017_563,1.0,2499,101,4,0,2,13,6,608,1,156,6,3,17,72,4,5,6,17,72,4,5,6,0,0,0,0,0,40,47,608,6,5,2,12,3,2,12,2
20180917,2118839820,2118913421,2118913421_1537129245_708,0.0,4935,101,1,6,2,23,0,764,1,764,0,3,4,7,2,4,7,4,7,2,4,7,0,0,0,0,0,1,1,764,0,5,3,10,2,3,10,3,10,2,
20180917,2118839820,2118913421,2118913421_1537129245_708,1.0,4935,101,1,6,2,23,0,764,1,764,0,3,4,7,2,4,7,4,7,2,4,7,0,0,0,0,0,1,1,764,0,5,3,10,2,3,10,3,10,2,
20180917,2118919566,2118885262,2118885262_1537112272_809,0.0,3248,101,2,6,2,9,0,360,1,360,0,5,2,6,3,2,6,2,6,3,2,6,0,0,0,0,0,0,608,6,3,3,22,3,3,22,3,22,3,3,
20180917,2118919566,2118885262,2118885262_1537112272_809,0.0,3248,101,2,6,2,9,0,360,1,360,0,5,2,6,3,2,6,2,6,3,2,6,0,0,0,0,0,0,608,6,3,3,22,3,3,22,3,22,3,3,
20180917,2118919566,2118885262,2118885262_1537112272_809,1.0,3248,101,2,6,2,9,0,360,1,360,0,5,2,6,3,2,6,2,6,3,2,6,0,0,0,0,0,0,608,6,3,3,22,3,3,22,3,22,3,3,
20180917,2118919566,2118885262,2118885262_1537112272_809,1.0,3248,101,2,6,2,9,0,360,1,360,0,5,2,6,3,2,6,2,6,3,2,6,0,0,0,0,0,0,608,6,3,3,22,3,3,22,3,22,3,3,
20180917,2118883921,2118794078,2118794078_1536999180_363,1.0,0,201,1,6,2,5,6,608,1,608,6,4,2,13,4,2,13,2,13,4,2,13,0,0,0,0,0,0,0,458,0,2,7,137,0,7,137,7,137,
20180917,2118968827,2118794078,2118794078_1536996235_474,0.0,0,101,4,6,2,23,0,360,1,360,0,3,5,170,0,5,170,5,170,0,5,170,0,0,0,0,0,0,0,458,0,2,7,137,0,7,137,
20180917,2118968827,2118794078,2118794078_1536996235_474,1.0,0,201,4,6,2,23,0,360,1,360,0,3,5,170,0,5,170,5,170,0,5,170,0,0,0,0,0,0,0,458,0,2,7,137,0,7,137,
20180917,2106268871,2118715390,2118715390_1537114666_182,1.0,2053,101,1,6,2,16,6,608,1,608,6,3,12,53,0,11,51,12,53,0,11,51,0,0,0,0,0,0,99,110,608,6,3,8,48,0,8,
20180917,2106268871,2118715390,2118715390_1537114666_182,0.0,2053,101,1,6,2,16,6,608,1,608,6,3,12,53,0,11,51,12,53,0,11,51,0,0,0,0,0,0,99,110,608,6,3,8,48,0,8,
20180917,2118086327,2118704810,2118704810_1537016936_806,1.0,0,201,1,3,2,17,6,608,1,608,6,3,9,35,2,9,35,0,0,0,0,0,9,35,2,9,35,60,61,608,6,3,2,6,6,2,6,2,6,6,
20180917,2118466525,2118704810,2118704810_1537016936_806,0.0,0,101,1,6,2,12,6,608,1,608,6,3,5,23,3,5,23,5,23,3,5,23,0,0,0,0,0,19,19,608,6,3,2,6,6,2,6,2,6,6,
20180917,2118466525,2118704810,2118704810_1537016936_806,1.0,0,201,1,6,2,12,6,608,1,608,6,3,5,23,3,5,23,5,23,3,5,23,0,0,0,0,0,19,19,608,6,3,2,6,6,2,6,2,6,6,6,
```

- 1) 错位问题难以定位；
- 2) 特征debug困难；
- 3) 不定长的稀疏特征无法表达；





A network diagram consisting of several blue circular nodes connected by thin blue lines, forming a web-like structure across the top half of the image.

# THANKS



A large, light gray watermark text "ITPUGB.NET" is oriented diagonally across the center of the image, partially overlapping the "THANKS" text.



Abstract geometric shapes in the bottom right corner, including overlapping triangles and curved bands in shades of pink, orange, yellow, and light blue.