阿里巴巴集团 旗下公司

天猫 TMALL.COM

阿里巴巴大数据竞赛

-天猫推荐大挑战技术交流

天猫-推荐算法团队

2014年3月

提纲

- 大赛背景&赛题简介
- 内部品牌推荐算法介绍
- 交流时间

天猫推荐总体情况

推荐产品

超过**40+**推荐产品

每天服务 **10M+**用户, 双11当天服务**36M+**用户

推荐算法

User2Items | Item2Items | Personalized Ranking | Others

推荐实体

品牌推荐

商品推荐

促销活动

等等

服务平台

Mobile

PC

天猫推荐总体情况-架构


Detail 给我推荐
天猫/淘宝 首页 logo 推荐
会员俱乐部推荐
图书城推荐
前端
匹配层
装配（参数化）
Ranker（CTR 预估）
检索
用户实时意图计算
在线实验框架
离线实验框架
推荐实体索引
Hadoop集群
ODPS
MPI集群
Spark集群

It’s a real problem

天猫品牌推荐业务

品牌是联接商家与消费者的纽带

日常：每天曝光用户超过4千万

2013.11.11， 接近6千万用户浏览会场

双11-会场个性化

每日9点
大牌上新！
15天全网
新品最低价
WHAT' S
NEW?

提前预知
明日大牌上新
下载天猫客户端
可提前0点购买哦!
COMING
SOON
我想看>

巨式国际
GIVH SHYH/巨式国际
上新34款 限时包邮 上新21件

粉红大布娃娃
粉红大布娃娃
春装上新5折优惠 上新7件

DIECHEN
蝶臣
春季上新全场3折送券 上新22件

妖精的口袋
ELF SACK
妖精的口袋
新品6.8折起 上新32件


春上新全店5折包邮


夏装新品4折起


新品立减20全场包邮

春装新品2折起包邮

春夏新品3折起体验价


新品九折包邮


连衣裙上新5折


莲灿高端新品上新

大赛背景与赛题简介

$$prob\_buy(brand_j|user_i, context)$$

5.7亿用户行为数据

比赛题目

开放数据	天猫用户在某一年04月-08月的品牌行为数据： 对品牌的点击、购买、收藏、加入购物车等。
预测数据	同样这些用户在同年9月份将会购买的品牌
评价指标	$Precision = \frac{\sum_i^N hitBrands_i}{\sum_i^N pBrands_i}$ $Recall = \frac{\sum_i^M hitBrands_i}{\sum_i^M bBrands_i}$ $F_1 = \frac{2 * P * R}{P + R}$

开放的数据

字段	字段说明	提取说明
user_id	用户标记	抽样&字段加密
time	行为时间	精度到天级别
action_type	用户对品牌的行为类型	包括点击、购买、加入购物车等。
brand_id	品牌ID	抽样&字段加密

提纲

- 大赛背景&赛题简介
- 内部品牌推荐算法介绍
- 交流时间

问题的定义

有些人尝试把问题转化为**评分预测**问题。

Type	Score
Click	1
Fav	2
Add2Cart	3
Buy	5

	$Brand_1$	$Brand_2$	……	$Brand_n$
$User_1$	1	?	…	?
$User_2$	?	5	…	3
$User_3$	3	?	…	?
…			…	
$User_m$				

SVD++　　Factorization Machines

问题的定义

有些人尝试把问题转化为**分类**问题。

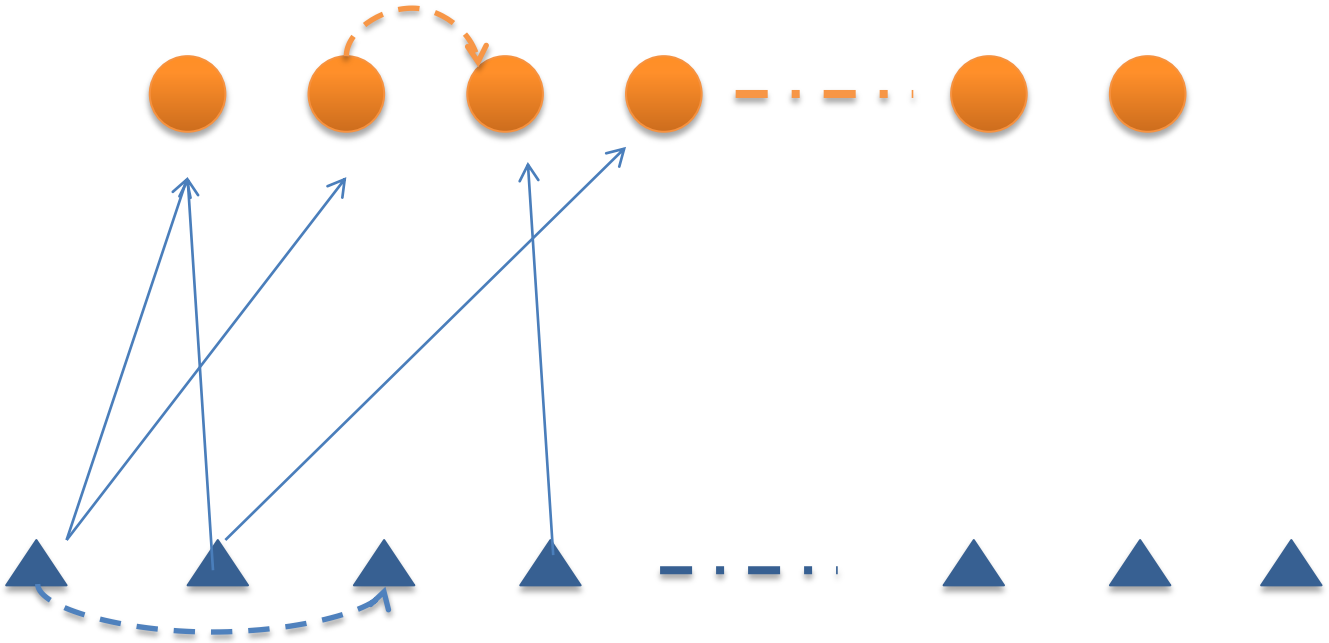
	f_1	f_2	…	f_n
…			…	
$User_i$-$Brand_j$	1	0	…	2
…			…	

Type	Label
Click	1
Fav	2
Add2Cart	3
Buy	4

各种分类算法…

问题的定义

有些人尝试把问题转化为**Graph**问题。



复杂网络（标签扩散、热传导）、...

问题的定义

在天猫，我们尝试把它转化为一个点击率预估问题

Training Stage

	f_1	f_2	…	f_n	if-click
…			…		0
$User_i$-$Brand_j$	1	0	…	2	1
…			…		…

Prediction Stage

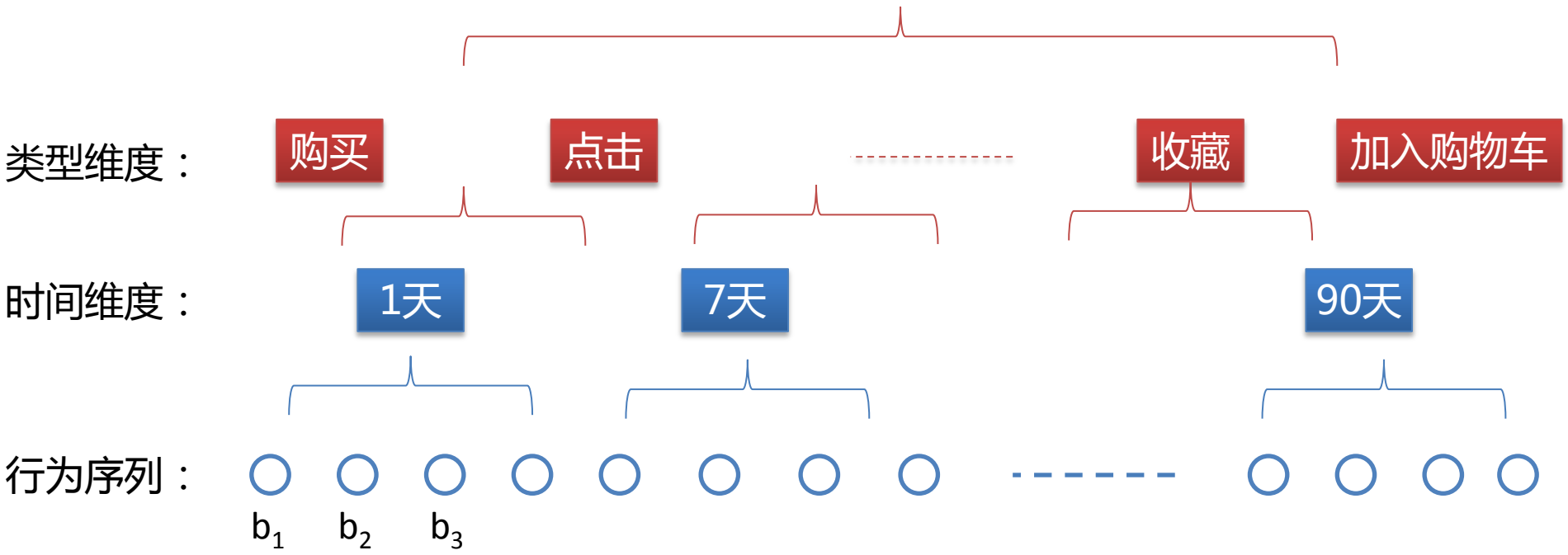
	f_1	f_2	…	f_n	Click prob
$User_i$-$Brand_x$	1	1	…	0	?

not the best, not the worst

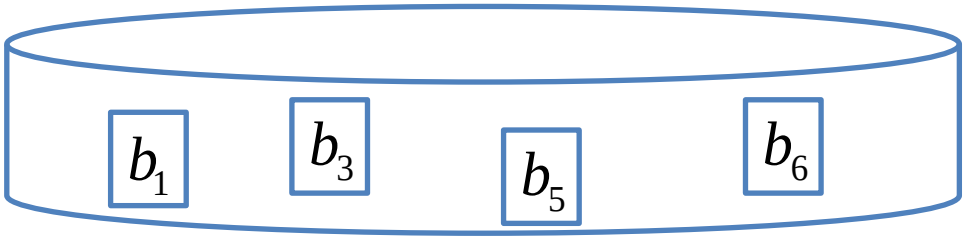
Feature Engineering

Feature Space：

	f_1	f_2	…	f_n	Click prob
$User_i$-$Brand_x$	1	1	…	0	?



偏好品牌集合
for user x



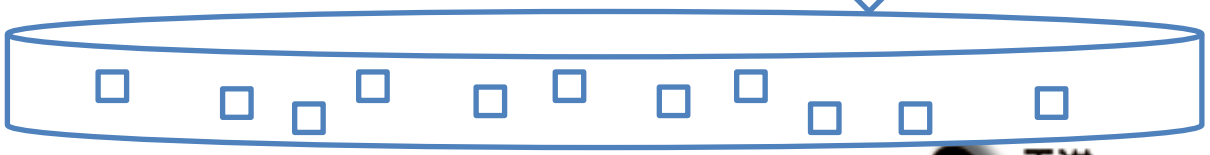
相似品牌

ItemBaseCF算法：

	b_i
b_i	sim

主品牌	相似品牌
b_1	b_{11} ; b_{12} ; b_{13} ; … ; b_{1h}
b_3	b_{21} ; b_{22} ; b_{23} ; … ; b_{2k}
b_5	……
b_6	b_{61} ; b_{62} ; b_{63} ; … ; b_{6k}

新品牌
for user x



Models

线性Model ： Logistic Regression

非线性Model ： Random Forest & GBDT

离线评估系统

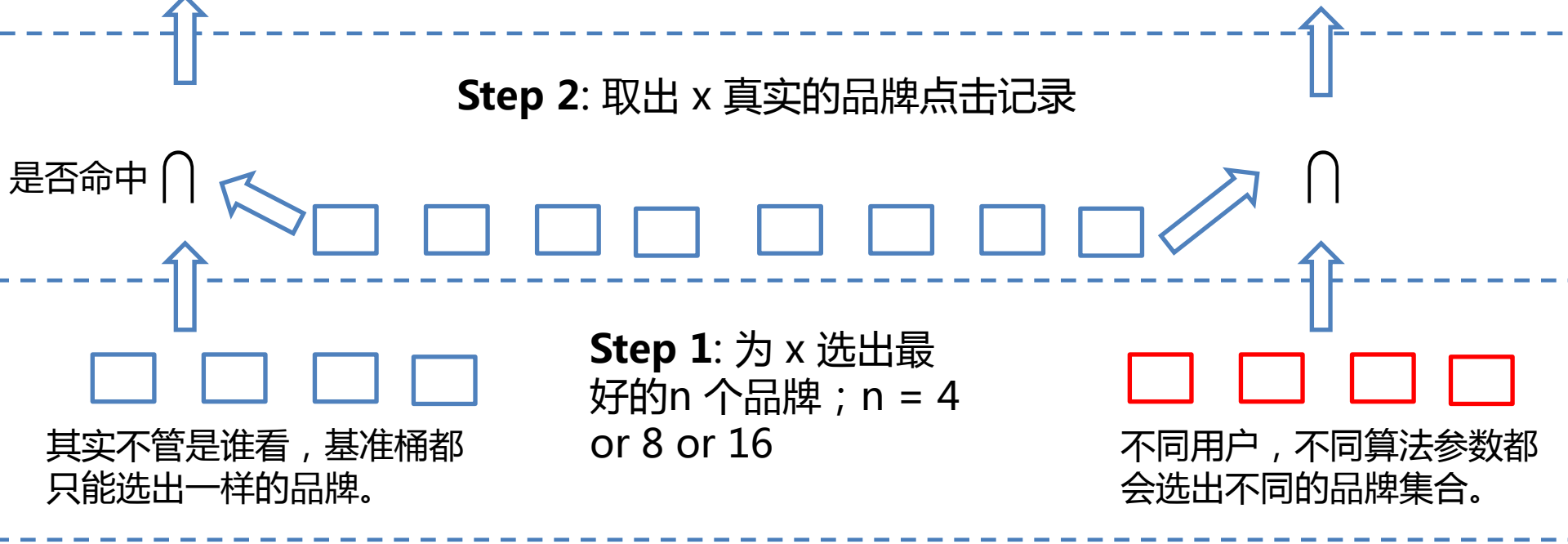
Step 4: 计算所有用户的命中率

$$hit\_rate = \frac{\sum_i hits(user_i)(总命中数)}{\sum_i delivers(user_i)(总曝光数)}$$

Step 3:计算单个用户**命中数**和**投放数**

for user x PK

Step 2: 取出 x 真实的品牌点击记录



Step 1: 为 x 选出最好的n 个品牌；n = 4 or 8 or 16

其实不管是谁看，基准桶都只能选出一样的品牌。

不同用户，不同算法参数都会选出不同的品牌集合。

基准桶

优化桶

品牌推荐算法的重要性

2013.11.11，近6千万用户访问会场

2013.08.28，我们做到了50%的提升

2013.11.11，我们做到了17%的提升

2014.11.11，你能做到了？%的提升

Welcome on board!

交流环节

竞赛互动平台：

1. **官方BBS：阿里云论坛**

2. **来往扎堆：数据魔法学院**

谢 谢！