

# “达观杯” 个性化推荐算法 挑战赛

团 队

提莫队长正在待命

团队成员

何从庆 曾露



# 团队成员

何从庆 湖南大学硕士 主要研究方向：数据挖掘、机器学习

曾露 华东理工大学硕士 主要研究方向：数据挖掘、自然语言处理



# 目录

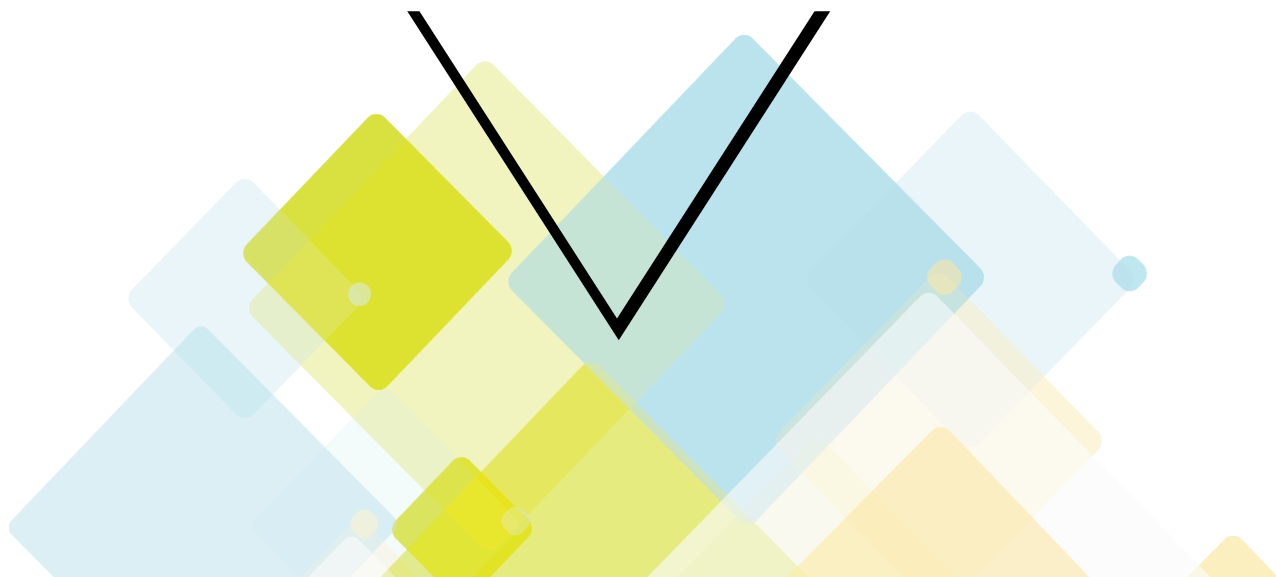
CONTENTS

- 1** ..... 赛题解读
- 2** ..... 算法思路
- 3** ..... 难点亮点
- 4** ..... 比赛总结



01

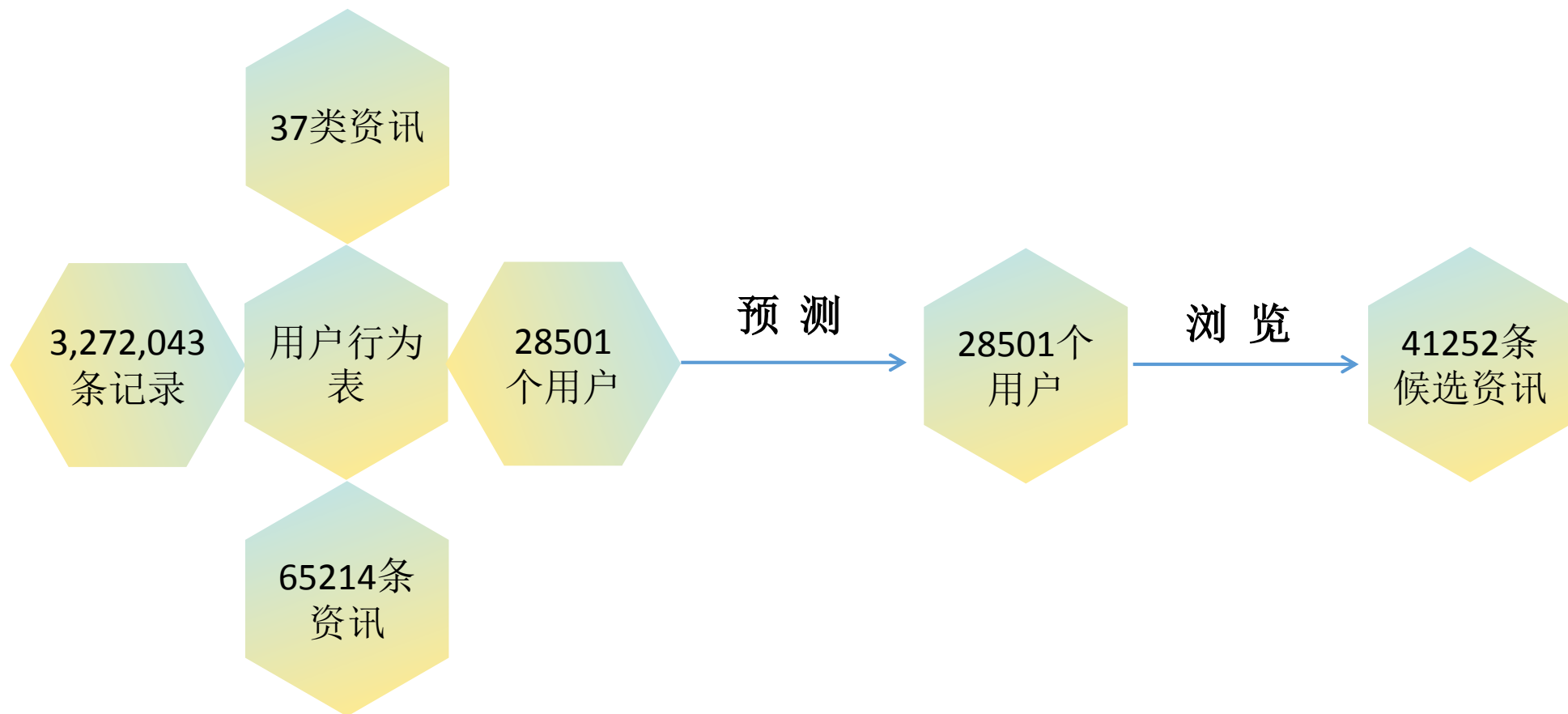
赛题解读



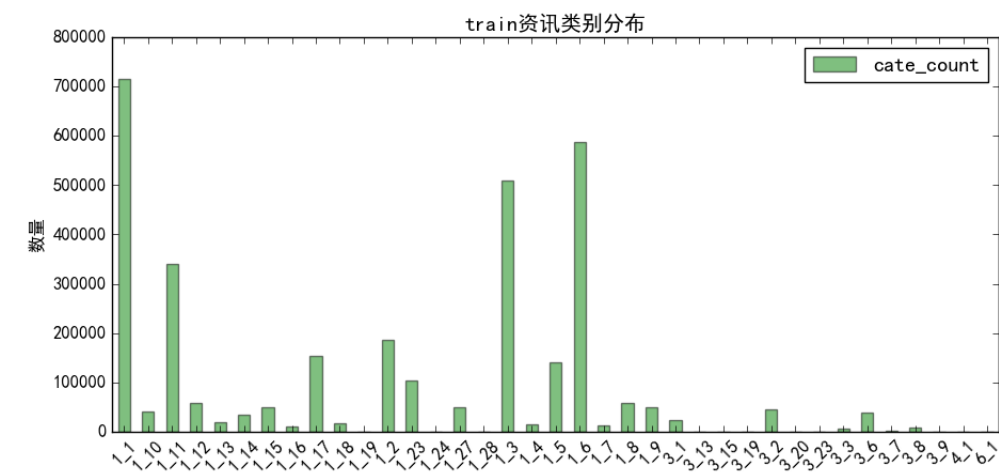
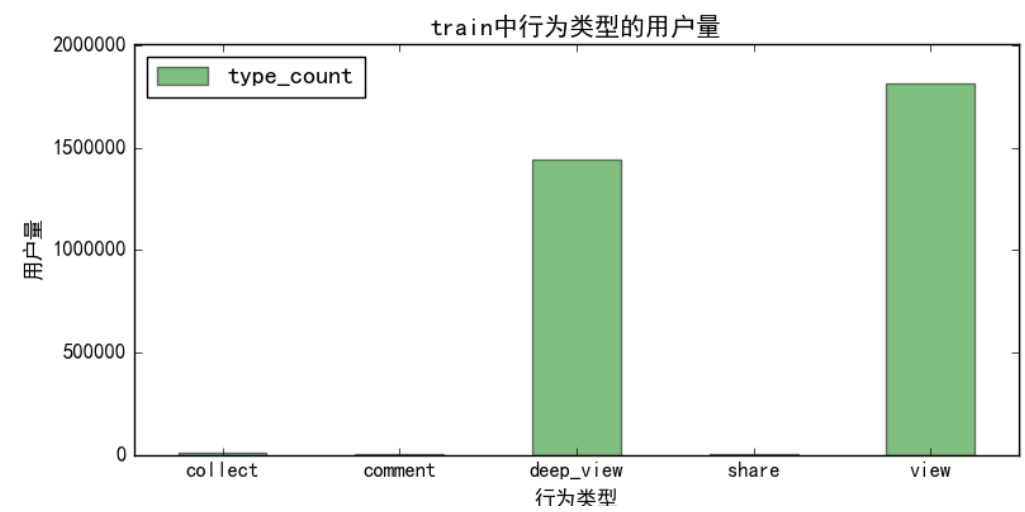
# 赛题与数据


## 数据

本赛题提供在2017年2月16日至2017年2月18日之间用户浏览资讯行为，预测在2017年2月19日用户最可能的5条浏览资讯内容。



# 数据分布





02

算法思路



## 2.1 机器学习的方法

### 训练集、测试集的构成

- 训练集的构成：考虑到数据的时效性，取最近一天的数据（2017年2月18日）作为训练集的正样本。同时，需要构造负样本，取最后一天的所有用户id, 选择当天各品类下的最热门的Top资讯，从而构成负样本。
- 测试集的构成：用candidate表中所有候选用户id以及test表中各品类下的最热门的top资讯数据集，构造测试样本。
- 采用常规的分类机器学习思想训练。
- 出现的问题：训练的模型达不到预期的效果，放弃！



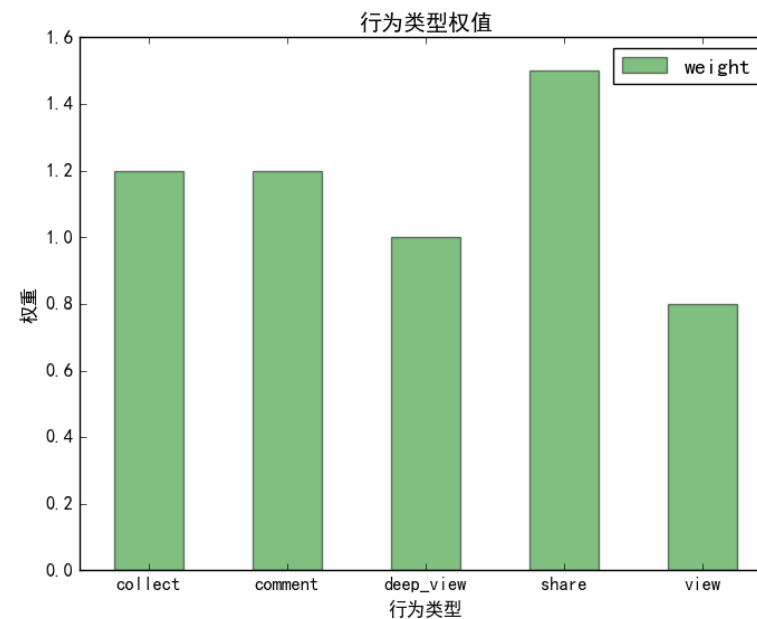
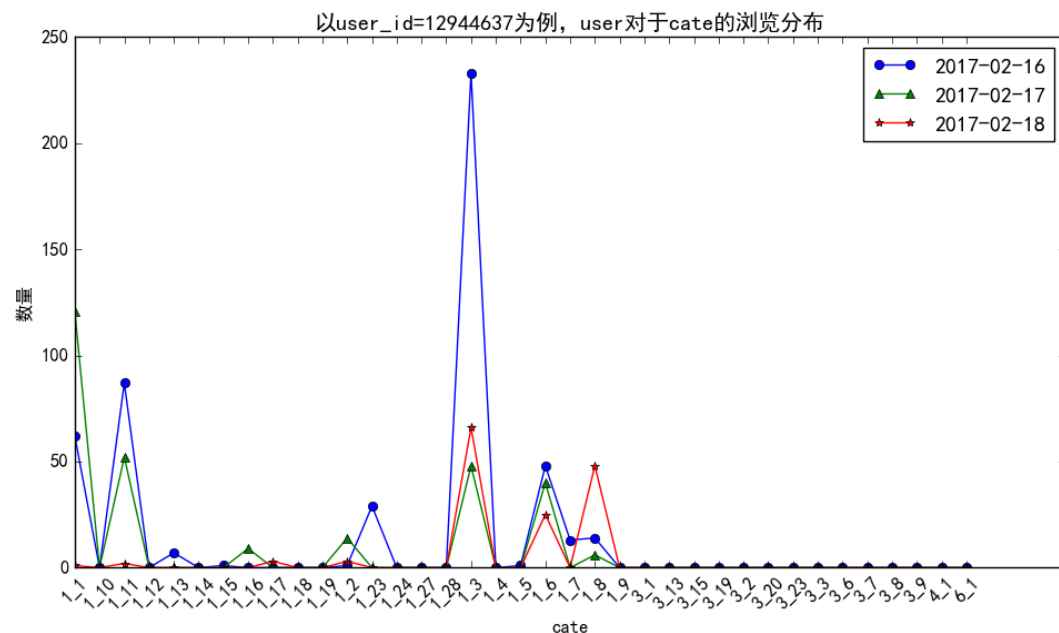
## 2.2 规则的方法

- 用户对于资讯类别的偏爱程度
- 资讯的热门程度

用户对于资讯的喜爱概率=用户对于资讯类别的偏爱程度\*资讯的热门程度

如果仅仅根据用户浏览资讯的行为进行计算用户的真实偏好，无法真实了解用户的偏好。会受到时间、行为类型的影响。

## 2.3 时间衰减、行为类型 优化——用户真实偏好



时间衰减函数:

$$V(T) = \frac{1}{1 + \alpha(T - t)} (\alpha \text{ 是衰减因子})$$

行为衰减:

$$W(action) = action\_type * weight$$

## 2.4 贝叶斯框架优化——用户真实偏好

(1) 计算一天用户对类别 $c_i$ 新闻的点击概率：

$$interest^t(cate = c_i) = p^t(click | cate = c_i) = \frac{p^t(cate = c_i | click) p^t(click)}{p^t(cate = c_i)}$$

其中， $p^t(cate = c_i | click) = \frac{N_i^t}{N_{total}^t}$  代表用户点击属于类别 $c_i$ 的概率，

$p^t(cate = c_i)$  表示资讯为 $c_i$ 的先验概率， $p^t(click)$  是用户点击的先验概率

## 2.4 贝叶斯框架优化——用户真实偏好

(2) 用户对类别  $c_i$  新闻的真实点击概率：

$$\begin{aligned} interest(cate = c_i) &= p(click | cate = c_i) = \frac{\sum_t N^t * interest^t(cate = c_i)}{\sum_t N^t} \\ &= \frac{\sum_t N^t * \frac{p^t(cate = c_i | click) p^t(click)}{p^t(cate = c_i)}}{\sum_t N^t} = \frac{\sum_t N^t * \frac{p^t(cate = c_i | click) p^t(click)}{p^t(cate = c_i)} + G}{\sum_t N^t + G} \end{aligned}$$

其中， $G$  是为了加入平滑，这种做法的好处是，如果用户的点击率较低，系统将会根据当前的新闻趋势进行推荐。同时用户的喜好也能够不断更新。

## 2.5 资讯热门度



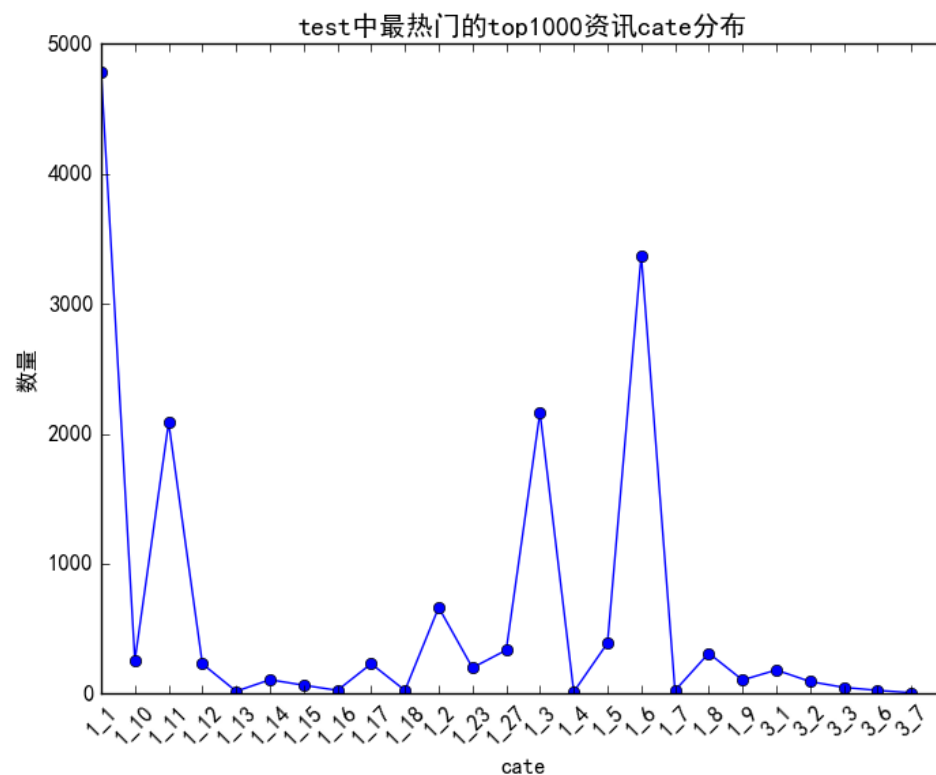
时间衰减函数: 
$$W(T) = \frac{1}{1 + \beta(T - t)}$$
 ( $\beta$ 是衰减因子)


## 2.5 测试集的构成

- candidate表的user
- test表中的各 cate 下的top30items

样本过滤:

- 用户已经产生过行为物品
- 候选物品以外的物品
- 用户未交互候选资讯





03

难点亮点



## 难点亮点

- 测试集的构造。（如果仅仅选择所有资讯下最热门的资讯，忽视了部分类别资讯，使得测试集构造不完整）
- 时间衰减函数的使用，有效的减小了随着时间变化对于用户偏好的影响，行为类别的加权，有效的减小各个行为对于用户偏好的影响。
- 贝叶斯框架的使用，借鉴 google 《personalized news recommendation based on click behavior》这篇文章，根据贝叶斯网络预测用户的资讯喜好。





04

比 赛 总 结



## 本次比赛的一些小遗憾

- 测试集资讯样本未能扩充，类似于时间衰减，将每日少量热门资讯加入测试集
- 对于时间的划分，不够细腻，因为资讯发布的时间不同，重要性也就不同
- 数据量较小，无法真实的反应用户的真实偏好，如果从一个较长的时间角度，可以更好的拟合用户的偏好



# THANKS

汇报人

何从庆