



# 推荐算法：如何打造和使用用户特征

王尧

# 个人介绍

---

- 王尧
- 在爱奇艺做过一段时间的算法工程师
- 当前任教于中原工学院软件学院

# 过去10年网站是如何变迁的？



# 过去10年网站是如何变迁的？

图片

专题

热点



宫斗剧厚马



刘天运学可



车上的都市风景

- 魏则西早年于墓门外 人在北京通国位前的前途去了
- 现在的潮人真让人头疼 将生儿事件还不知显示的时代
- 周小 给谁个好评 杆的留后 能投就是天也出是感本
- 吴梦瑶还有多久能和六品 出 越多操得越快？

新闻

看河南

2017-11-28

- 习近平会见缅甸国防军总司令
- 习近平与中共与世界共产党高层对话会开幕式
- 这些“泥瓦匠”“喷漆工”为何能高达中南海？
- 以筑新征程 墨成孩子的“未来史拉”
- 本周除重下台说 经济还舒服一个不一路行幸
- 伏虎史：近期中共对香港的这内以举动 颇有意味
- “老康生”的党与政：无发信言兵队及潜藏行为
- 津市市三新总流上任中方是否发发电？外交部的回应
- 毛岸英牺牲50周年：毛泽东曾称其遇害30年
- 中国与美国批准经济合作与贸易便利化新议定书万可能
- 天津落马正厅即将被听 此前本公 其被双开消息
- 魏德福指迷中必成：总理以同能同在中五人被上驾
- 北京花100亿办了一个大子 领导无子考核都出名之
- 成功入选《世界记忆名录》的叶青文 你认识几个？
- 处级警官英老举根和级级警官：对核心证人利用保护
- 村寨角成老人住家新空：面对朋友好 反目成老仇
- 杭州桐庐小学生放寒假比放寒假片
- 广东男子“人接3000条骚扰电话 到因为拉黑”人

视频

秒拍

综艺



通感二已成功发射



张西波武学奇史

- 中国“王胡安”号深能大取已超元
- 中国共商标准 大帝制得管哥
- 心为中安抗并 门企和解能快
- 雨中人形空竹 街舞六元六线
- 1-1 通感二白竹的景
- 陈月先道在传与风波
- 风骚操作！吕超如大超才晋点
- 七球看台竟又重现越战场面

# 过去10年网站是如何变迁的？

---



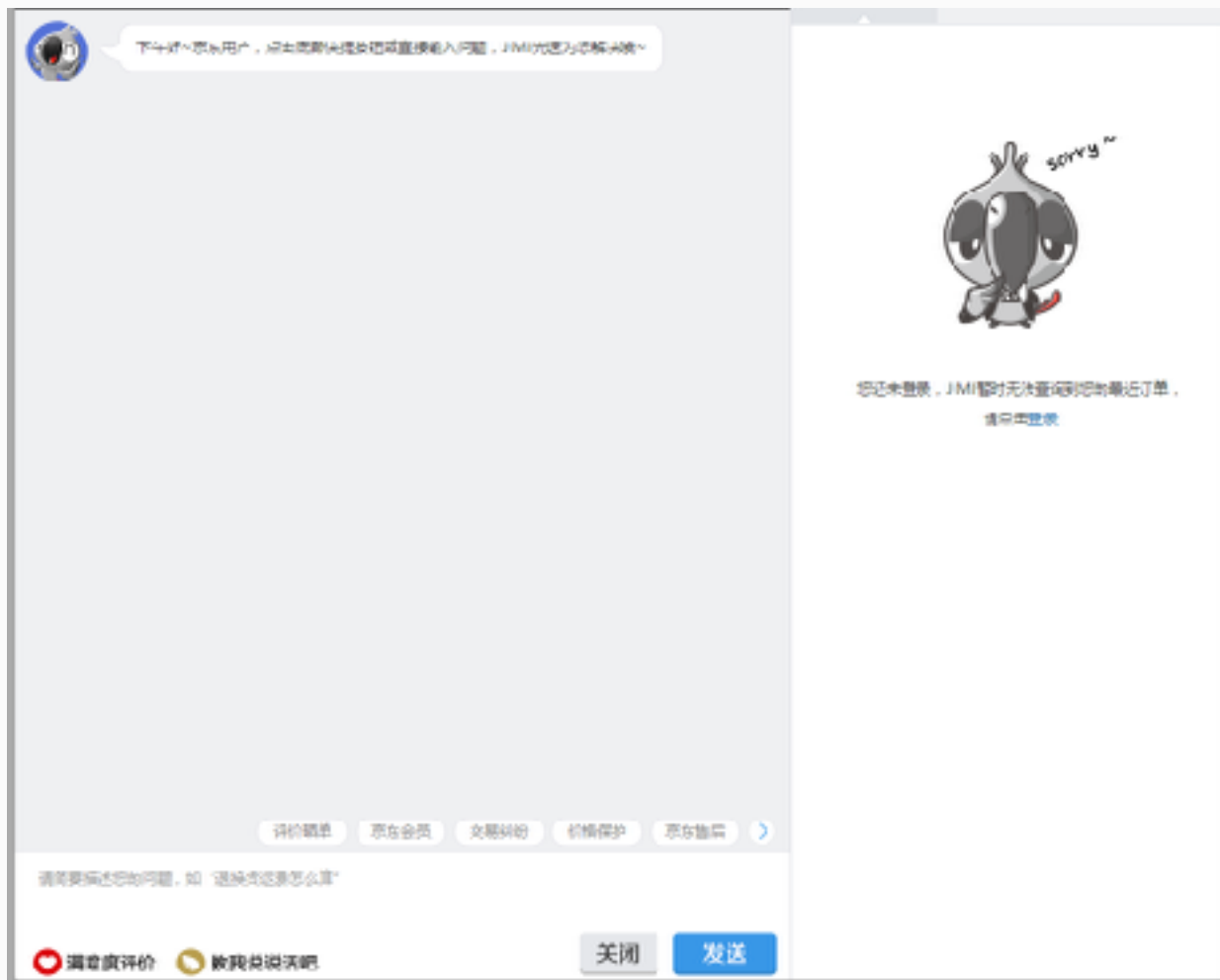
# 过去10年网站是如何变迁的？



# 过去10年网站是如何变迁的？



# 过去10年网站是如何变迁的？





# 网站的变迁

---

- 网站的变迁反映了互联网人对于互联网理解的发展
- 信息发布：网站黄页、新闻平台
- 信息处理：搜索引擎
- 讨论：随意或认真发言的各类社区
- 认真听：聊天机器人

# 听和说哪个更困难？

---

- 说是在相同中寻找不同
- 听是在不同中寻找相同
- 为什么我们需要用户特征？
- 对于互联网应用来说，用户特征是不同中最重要的相同
- 如果我们需要对用户进行个性化的服务，必须在用户复杂的信息中找到能抽象为用户特征的信息

## 简单的用户特征是什么？

---

```
String id;           // 用户id
String userName;     // 用户姓名
Date birthday;       // 用户生日
Integer age;          // 用户年龄
String password;      // 用户密码
String email;         // 用户邮箱
String phoneNum;      // 用户手机号
```

## 如何使用用户的简单特征来推荐？

---

- 如果用户年龄在某个区间内，可以推荐...
- 如果用户的手机号是某个地域的手机号，可以推荐...
- 如果到了用户生日，可以推荐...

## 比这些复杂的？

---

- 带有一些挖掘出来的信息，如：
  - 用户推测年龄
  - 用户推测性别
  - 用户推测城市
  - 用户活跃度
  - 用户使用APP时间偏好
  - 用户对APP的频道偏好
  - 用户忠诚度
  - 用户对APP的使用时长偏好
  - 用户对APP的平台偏好
  - 用户的新鲜度偏好
  - ...

# 如何使用挖掘特征来推荐？

- 我们设定的描述维度越多，处理就越复杂：
  - 多个维度如何共同起作用
  - 当维度数据缺失时如何应对
- 产品的特征生成：
  - 统计购买该产品的用户的属性加和求平均
- 判断是否推荐某个产品给用户：
  - LR：逻辑回归Logistic Regression
    - $$g(x) = \frac{1}{1 + e^{-\sum k_i x_i}}$$
    - 结合用户实际使用APP的数据，训练用户对产品的哪种特征更敏感

# 用户特征标签化

---

- 对于很多提供信息服务的网站来说，用户的交互往往和信息本身相关，如：
  - 新闻网站：用户钟爱的新闻
  - 视频网站：用户钟爱的视频的标题和描述信息
  - 音乐网站：用户钟爱的音乐演唱者和音乐所属分类
  - 购物网站：用户钟爱的商品的描述信息

# 如何使用用户标签来推荐？

---

- 产品特征生成：
  - 找一个小编去进行标签标注
  - 统计所有购买该产品的用户的标签的出现情况、排序后取topN
- 判断一个产品是否推荐给用户：
  - 直接进行标签匹配



# 标签匹配的问题

---

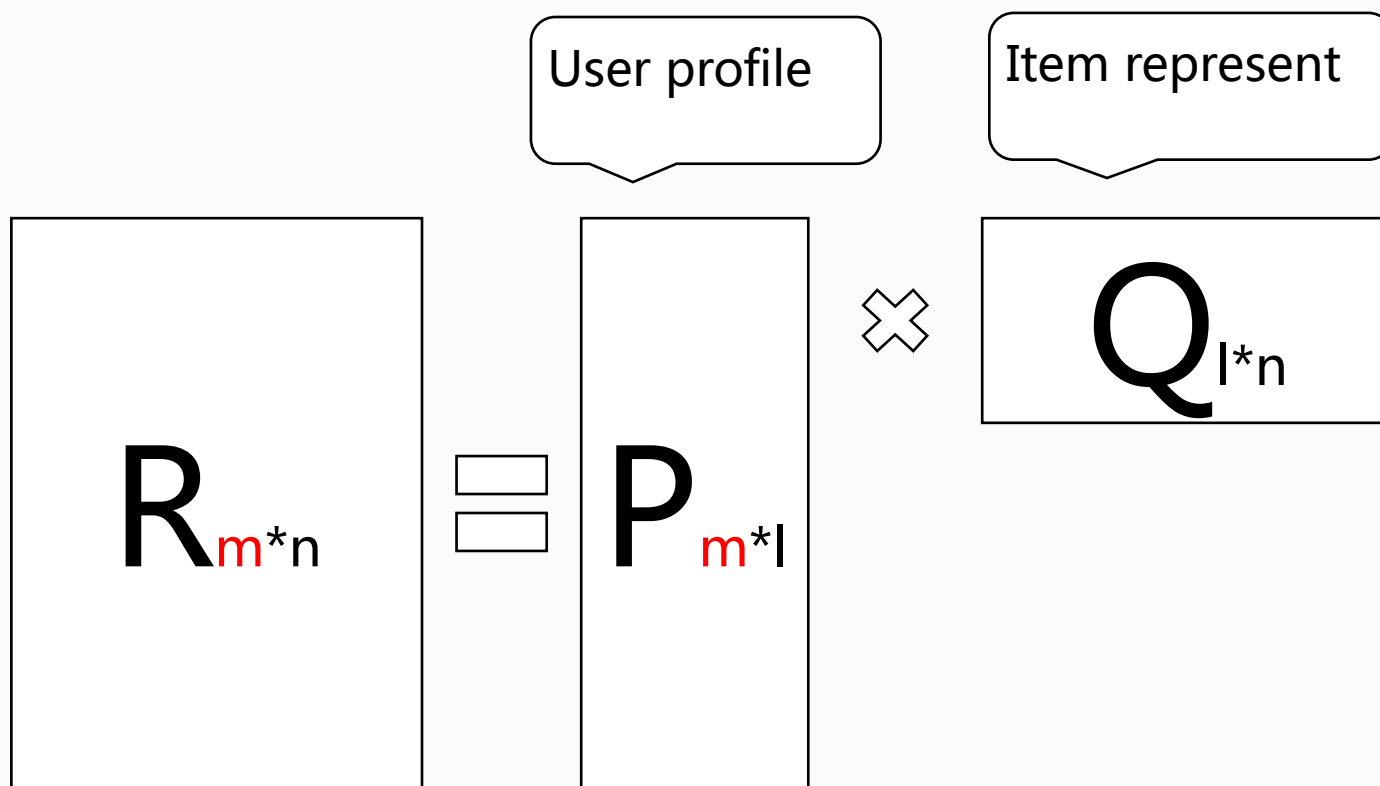
- 本身抽象为维度是为了在不同中找相同，现在反而出现更多不同的标签
- 维度数量不可控重合度低
- 无效标签浪费资源
- 标签相比固定特征而言，使用性较低

## 如何把《蒙娜丽莎》说给别人听？

- 使用隐语义：
  - 可描述特征+不可描述特征
  - 不可描述特征只有一个编号

	The Matrix	Titanic	Die Hard	Forrest Gump	Wall-E
John	5	1		2	2
Lucy	1	5	2	5	5
Eric	2	?	3	5	4
Diane	4	3	5	3	

# 矩阵分解



# 如何使用隐语义来推荐？

---

- 产品特征生成：
  - 通过机器学习训练出来
- 判断一个产品是否推荐给用户：
  - 通过特征维度的匹配

# Q&A

---

- Thank you