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？

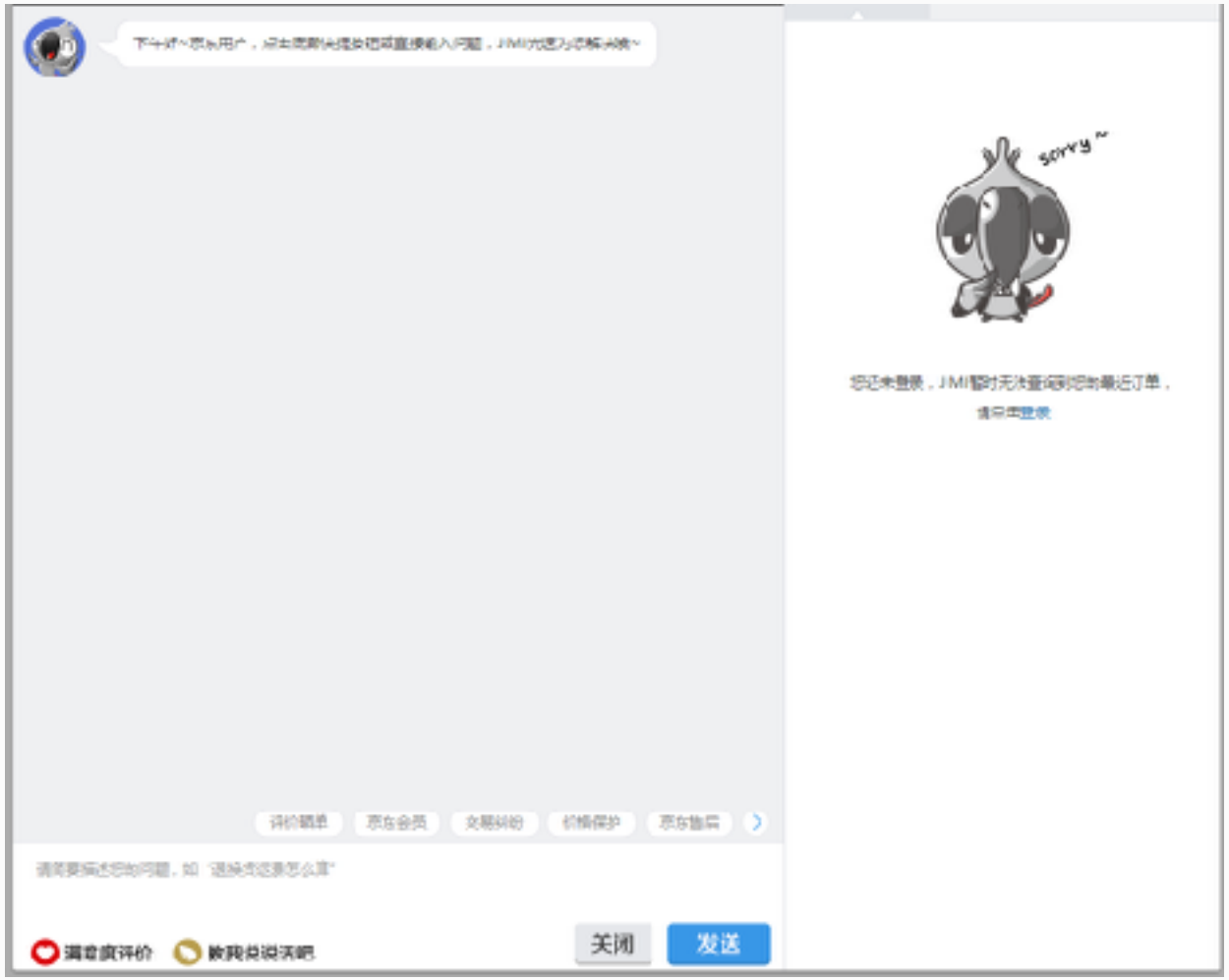
过去10年网站是如何变迁的？

GDG Zhengzhou

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？



GDG Zhengzhou

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

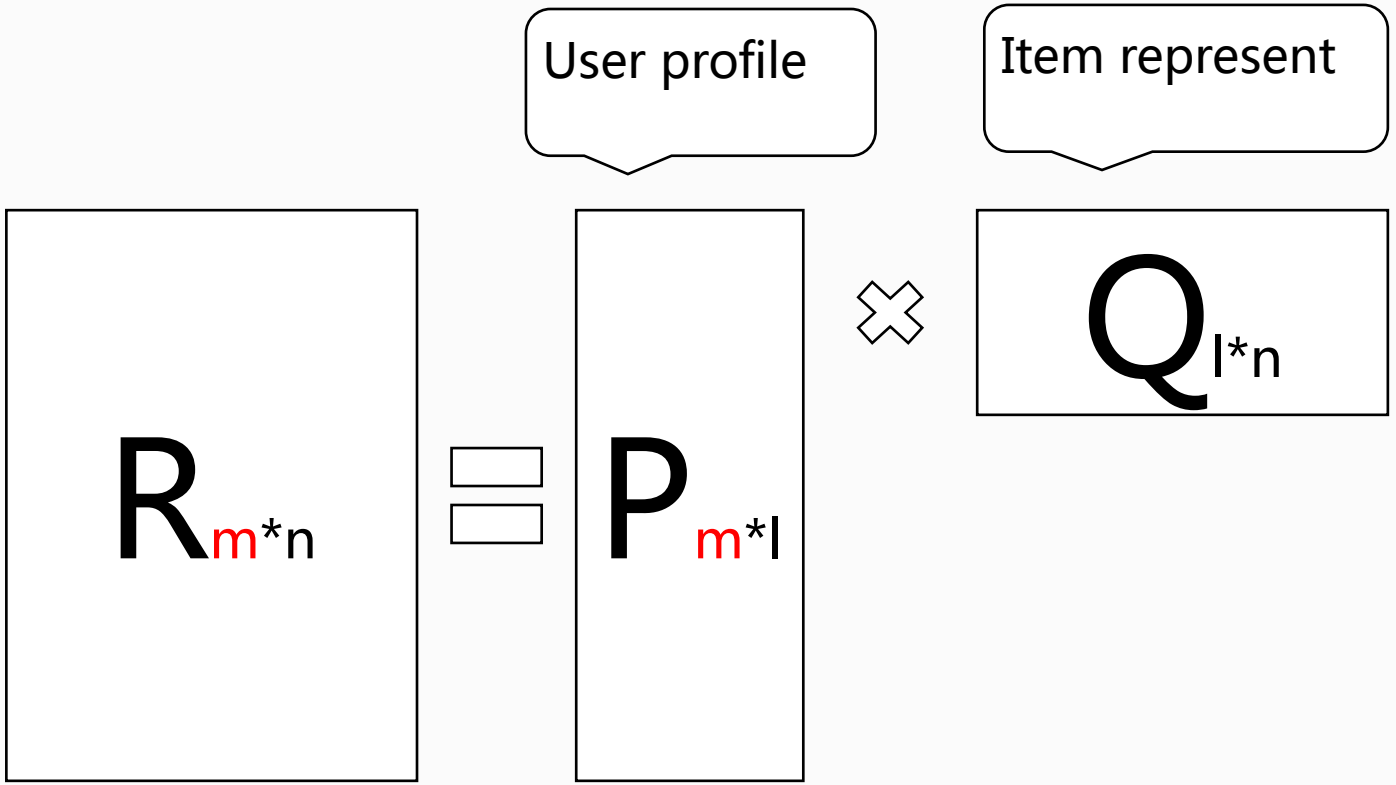
- 标签相比固定特征而言，使用性较低

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

GDG Zhengzhou