



# 数据挖掘导论

## Introduction to Data Mining

### 第六章 推荐系统

刘 淇

Email: [qiliuql@ustc.edu.cn](mailto:qiliuql@ustc.edu.cn)

课程主页:

<http://staff.ustc.edu.cn/~qiliuql/DM2017YZ.html>



# 主要目标

2

- 了解推荐系统的背景和现状
- 总结推荐算法的主要思想、技术方案
- 建立推荐系统与数据挖掘的关联
- 学习推荐算法的设计（研究）案例
- 探讨推荐算法的未来研究方向



# 主要内容

3

- 什么是推荐系统
  - 背景、定义、应用场景
- 推荐方法概述
  - 兴趣建模
  - 推荐算法设计
  - 推荐结果的评估
- 案例学习
  - 基于用户兴趣扩展的个性化推荐方法
  - 面向推荐系统的纠结心理挖掘
- 小结及未来的路
- 资料推荐



# 主要内容

4

## □ 什么是推荐系统

- 背景、定义、应用场景

## □ 推荐方法概述

- 兴趣建模
- 推荐算法设计
- 推荐结果的评估

## □ 案例学习

- 基于用户兴趣扩展的个性化推荐方法
- 面向推荐系统的纠结心理挖掘

## □ 小结及未来的路

## □ 资料推荐

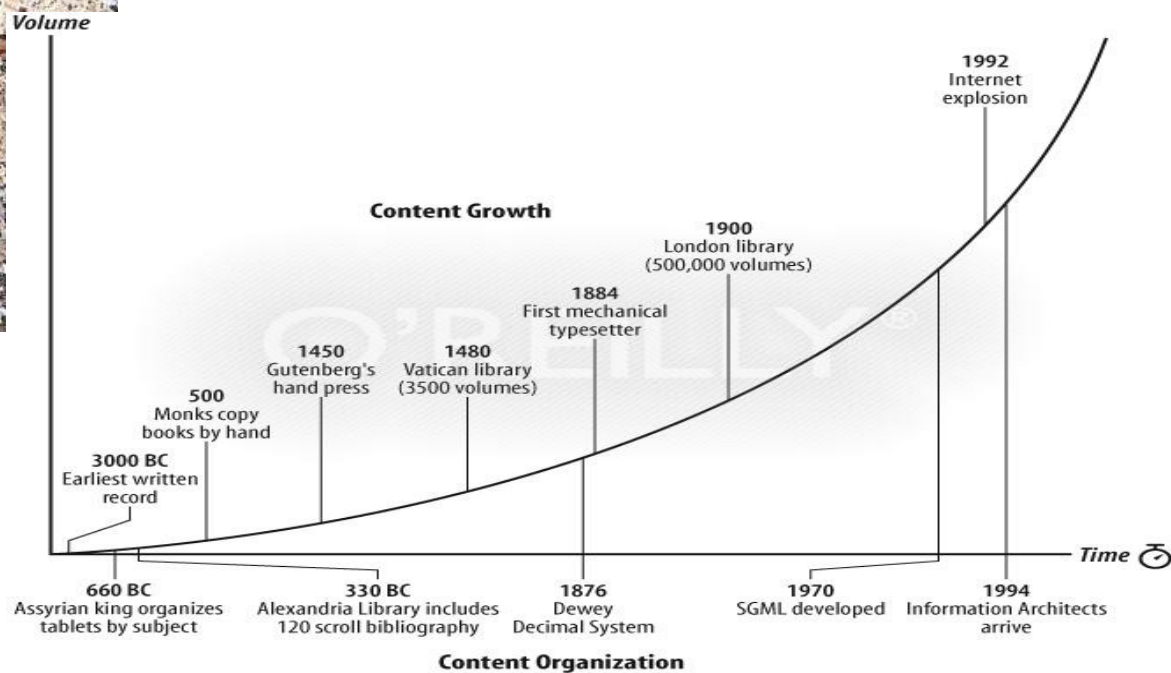


# 为什么要推荐

5



## Information overload



**We are leaving the age of information and entering the age of recommendation.**

**Chris Anderson in *The Long Tail***



# 背景

6

## 原因示例1



信息推荐!!

返回信息量太大







# 背景

7

## □ 原因示例-2



个性化信息推荐！！

08秋专柜款 JACK JONES 肩章二扣款修身小西服 货号(f141) 一口价 188.0元

08秋专柜款 JACK JONES 修身休闲小西服 货号(8862) 灰色 一口价 138.0元

08新 ZARA 单扣时尚修身麻料休闲西装 货号(183)米色 一口价 128.0元

皇冠 08秋JACK JONES立体翻领双排扣涂层短款风衣 货号(807) 一口价 178.0元

08新 Dior 经典双扣时尚修身英伦风格休闲西装 货号(181) 一口价 128.0元

08秋 JACK JONES 多袋涂层短款小风衣 货号(297) 一口价 178.0元

08秋专柜款 JACK JONES 休闲小西服 货号(8862) 米白 一口价 138.0元

ZARA 单扣时尚修身麻料休闲西装 货号(183) 一口价 128.0元



# 背景

8

## □ 推荐系统



信息技术



网络商家



项目



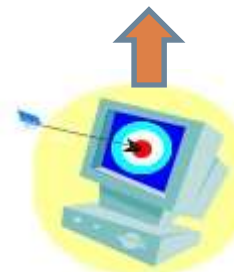
普通用户

推荐系统



- 项目过滤
- 主动服务

商家增加收益



用户各得所需





# 推荐无处不在—豆瓣

9

## I Am Legend



放在你的blog里!

增改描述、海报图片

简体中文名: 我是传奇

编剧: Mark Protosevich / Akiva Goldsman / Richard Matheson

导演: Francis Lawrence

主演: Will Smith / Alice Braga / Charlie Tahan

官方网站: <http://iamlegend.warnerbros.com/>

上映年度: 2007

语言: 英语

制片国家/地区: 美国

imdb链接: [tt0480249](http://tt0480249)

我看过这部电影 修改 删除

我的评价: ★★★★★ 力荐



推荐

喜欢看"这部电影"的人也喜欢



机械公敌



全民超人



国家宝藏2: 古籍秘辛



通缉令



科洛弗档案



300 / 300死士 / 300斯巴达勇士

Gerard Butler / Vincent Regan / Lena He

看过 想看 没兴趣



钢铁侠



心灵传输者



300



迷雾



国家公敌



Iron Man / 铁人 / 钢铁侠

Robert Downey Jr. / Terrence Howard / 钢铁侠 / Art Marcum / Matt Holloway / Ma

看过 想看 没兴趣



# 推荐无处不在—淘宝

10

我的淘宝



我浏览过的宝贝



1.00元



14.96元



5580.00元



5799.00元

猜你喜欢宝贝



【果园老农  
葡萄干

3.90元



三皇冠！月  
销三吨 新货

26.90元



【PC大佬】  
联想

6099.00元



包邮促销！  
月销2万斤

13.98元

商品页面



浏览了该宝贝的会员还浏览了



09最新包装 大兴安岭野生蓝莓  
果干 原味无糖保护视力 五袋  
包邮~

¥32.0元



【一斤包邮送礼】大兴安岭 纯  
野生黑蚂蚁 风湿病克星 养肝  
降血糖

¥5.0元



【五钻石信誉】特等野生枸杞  
补肾益精养肝明目 美颜佳  
品 特价

¥3.1元



# 推荐无处不在—Amazon

11

amazon.com

Hello. Sign in to get [personalized recommendations](#). New customer? [Start here](#).

Your Amazon.com [Today's Deals](#) [Gifts & Wish Lists](#) [Gift Cards](#)

[Shop All Departments](#)

[Books](#)  
[Books](#)  
[Kindle](#)  
[Textbooks](#)  
[Magazines & Newspapers](#)

[Movies, Music & Games](#)

[Computer Components](#)  
[Office Products & Supplies](#)

[Your Amazon.com](#)

- [Today's Recommendations For You](#)
- [Your Browsing History](#)
- [Rate These Items](#)
- [Improve Your Recommendations](#)

These recommendations are based on: [your most recently viewed items](#).

- 1. [The Time Paradox \(Artemis Fowl, Book 6\)](#)**  
by Eoin Colfer (Jul 15, 2008)  
Average Customer Review: [★★★★☆](#) (60)  
In Stock

**List Price:** \$47.99  
**Price:** \$10.79  
[63 used & new](#) from \$7.50

☐ I own it ☐ Not interested ☐ [Rate it](#)

Recommended because you recently viewed [Brisingr \(Inheritance, Book 3\)](#) ([Fix this](#))

[Add to cart](#) [Add to Wish List](#)
- 2. [Inkdeath \(Inkheart\)](#)**  
by Cornelia Funke (Sep 26, 2008)  
Average Customer Review: [★★★★★](#) (1)  
In Stock

**List Price:** \$24.99  
**Price:** \$14.99  
[7 used & new](#) from \$14.99

☐ I own it ☐ Not interested ☐ [Rate it](#)

Recommended because you recently viewed [Brisingr \(Inheritance, Book 3\)](#) ([Fix this](#))

[Add to cart](#) [Add to Wish List](#)

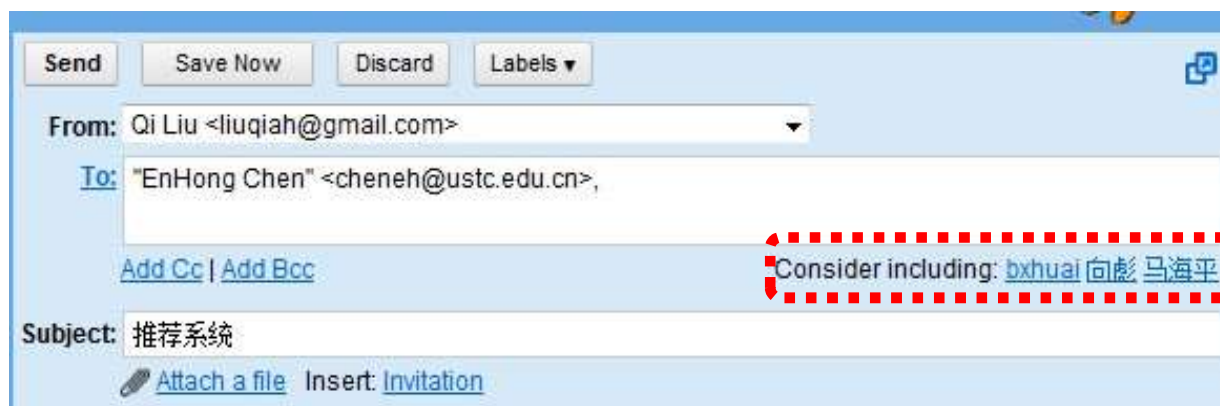




# 背景

12

## □ 推荐无处不在





# 面向推荐算法的竞赛层出不穷

13

## 首届全国大学生数据挖掘邀请赛



Will they contact?



# 面向推荐算法的竞赛层出不穷

14

**Netflix Prize**

Home Rules Leaderboard Update Download

## Leaderboard

Showing Test Score:

Display top:

**被评选为09年IT行业  
100件最重要大事之一**

Rank	Team Name	dev. test score	improvement	Best Submit Time
1	Recommender Systems Group	0.8507	10.0%	2008-07-28 18:18:28
2	The Algorithm	0.8507	10.0%	2008-07-28 18:38:22
3	Netflix Prize Team	0.8503	8.8%	2008-07-19 21:24:49
4	Netflix Prize and Recommender Systems Group	0.8500	8.8%	2008-07-18 11:12:21
5	Variable Reduction	0.8501	8.8%	2008-07-18 11:12:21
6	Practical Team	0.8504	8.7%	2008-06-24 12:05:56
7	Netflix Prize Team	0.8501	8.7%	2008-05-13 18:14:19
8	Rank...	0.8512	8.8%	2008-07-24 17:18:43







# 面向推荐算法的竞赛层出不穷

15



阿里移动推荐算法

已结束



2015/07/01

¥ 300000

7186

MobileDM&HuMoComp 2015: The First International Workshop on Mobile Data Mining & Human Mobility Computing(ICDM 2015)

TIANCHI天池

首页

天池大赛

数据实验室

天池科学家

互动

御膳房

## Organizers

### Workshop Chairs:

- Rong Jin (iDST, Alibaba Group)
- Yixin Chen (Washington University)
- Qi Liu (University of Science and Technology of China)
- Zhongyi Liu (Alibaba Group)
- Nicholas Jing Yuan (Microsoft Research Asia)
- Rui Zhang (University Of Melbourne)
- Kai Zheng (University Of Queensland)

## ICDM 2015

IEEE International Conference on Data Mining  
Atlantic City, NJ, USA. November 14-17, 2015





# 面向推荐算法的竞赛层出不穷

16



## 2016Byte Cup国际机器学习竞赛

主办方：IEEE中国 今日头条

报名时间：2016.08.05 00:00——2016.11.20 23:59

比赛时间：2016.08.15 00:00——2016.11.20 23:59



# 主要内容

17

- 什么是推荐系统
  - 背景、定义、应用场景
- 推荐方法概述
  - 兴趣建模
  - 推荐算法设计
  - 推荐结果的评估
- 案例学习
  - 基于用户兴趣扩展的个性化推荐方法
  - 面向推荐系统的纠结心理挖掘
- 小结及未来的路
- 资料推荐

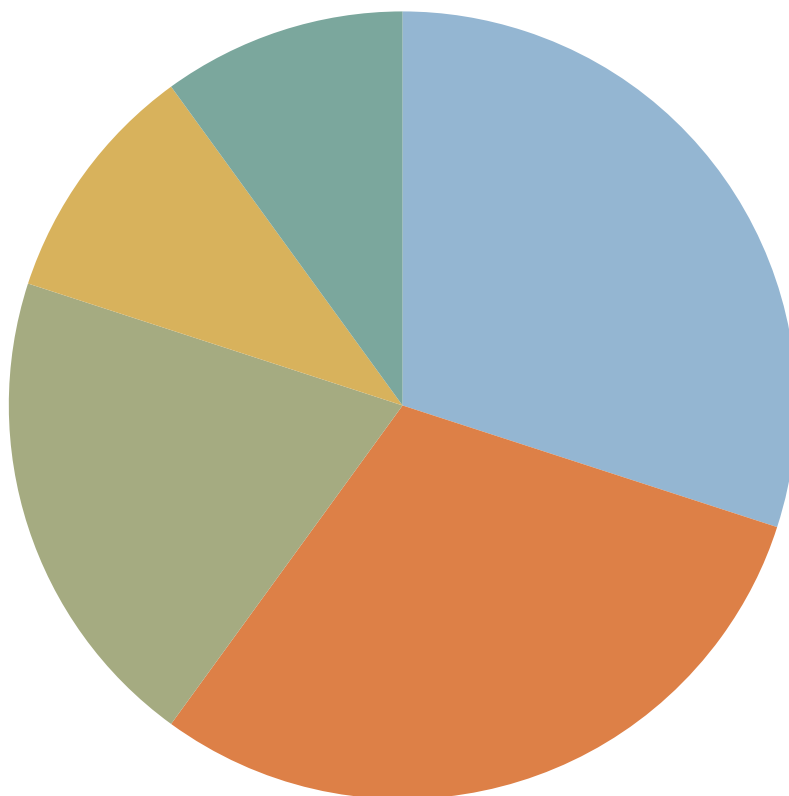


# 背景及相关工作

18

## □ 推荐系统

□ 首先是一个 “System”



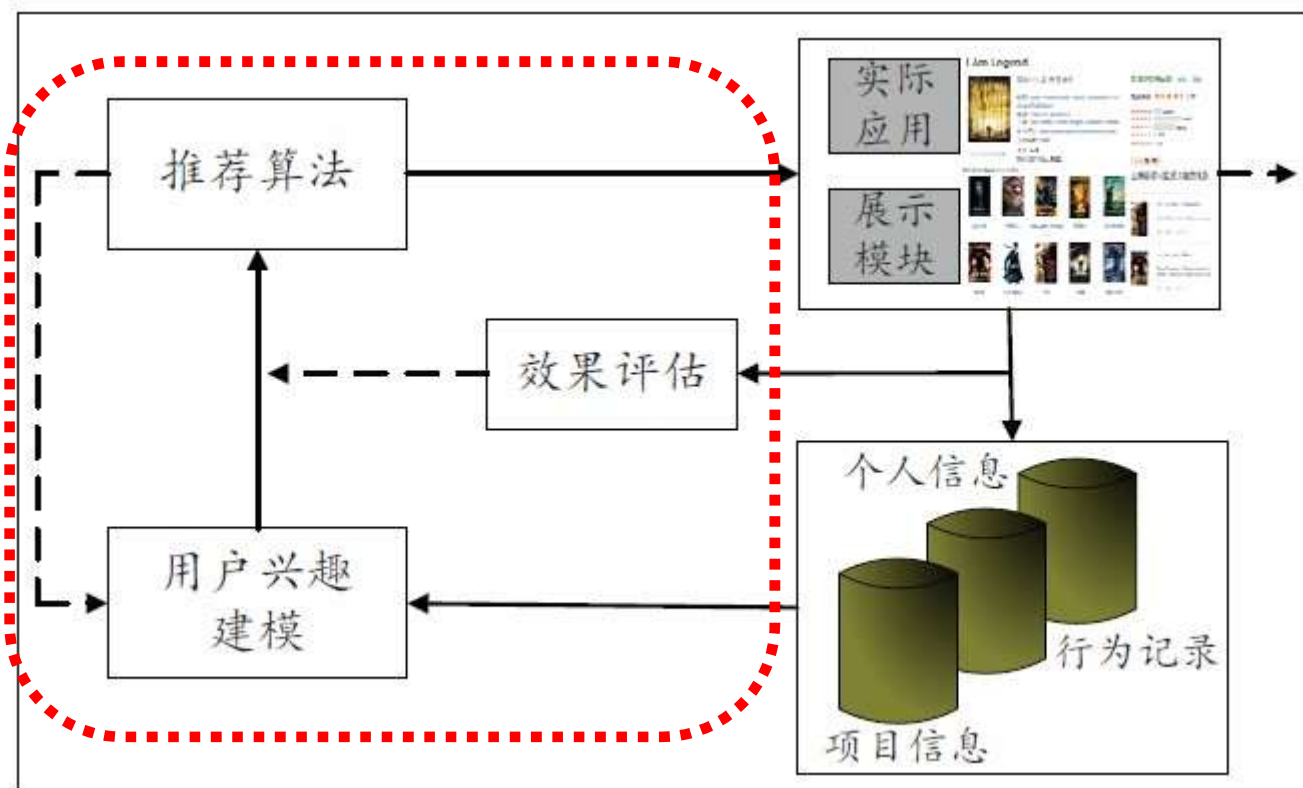
UI/UE	30%
Data	30%
Domain Knowledge	20%
Algorithm	10%
Others	10%



# 背景及相关工作

19

## □ 推荐系统框架





# 相关工作

20

## □ 推荐方法分类







# 主要内容

21

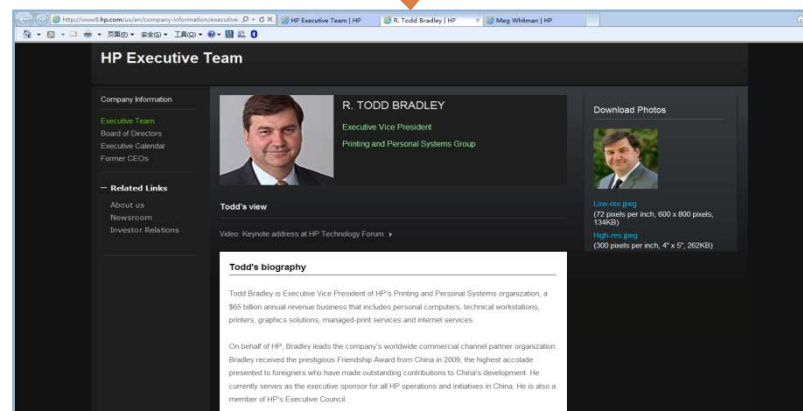
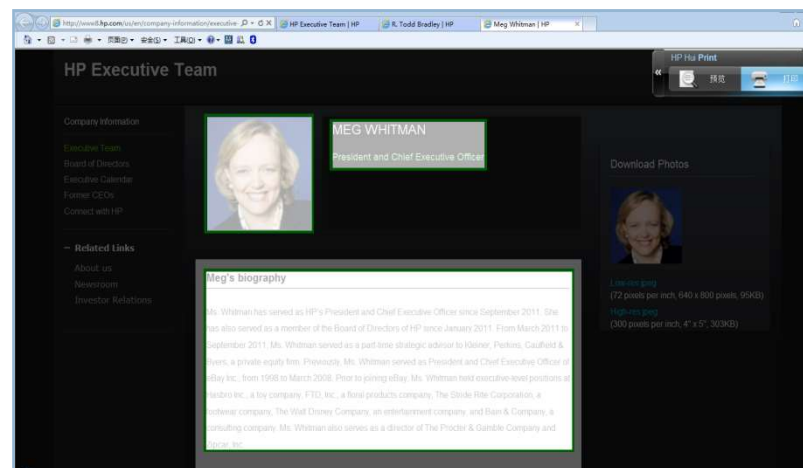
- 什么是推荐系统
  - 背景、定义、应用场景
- 推荐方法概述
  - 兴趣建模
  - 推荐算法设计
  - 推荐结果的评估
- 案例学习
  - 基于用户兴趣扩展的个性化推荐方法
  - 面向推荐系统的纠结心理挖掘
- 小结及未来的路
- 资料推荐



# 兴趣建模

22

- 非个性化兴趣模型
  - 无法识别单个用户的ID
  - 群体智慧

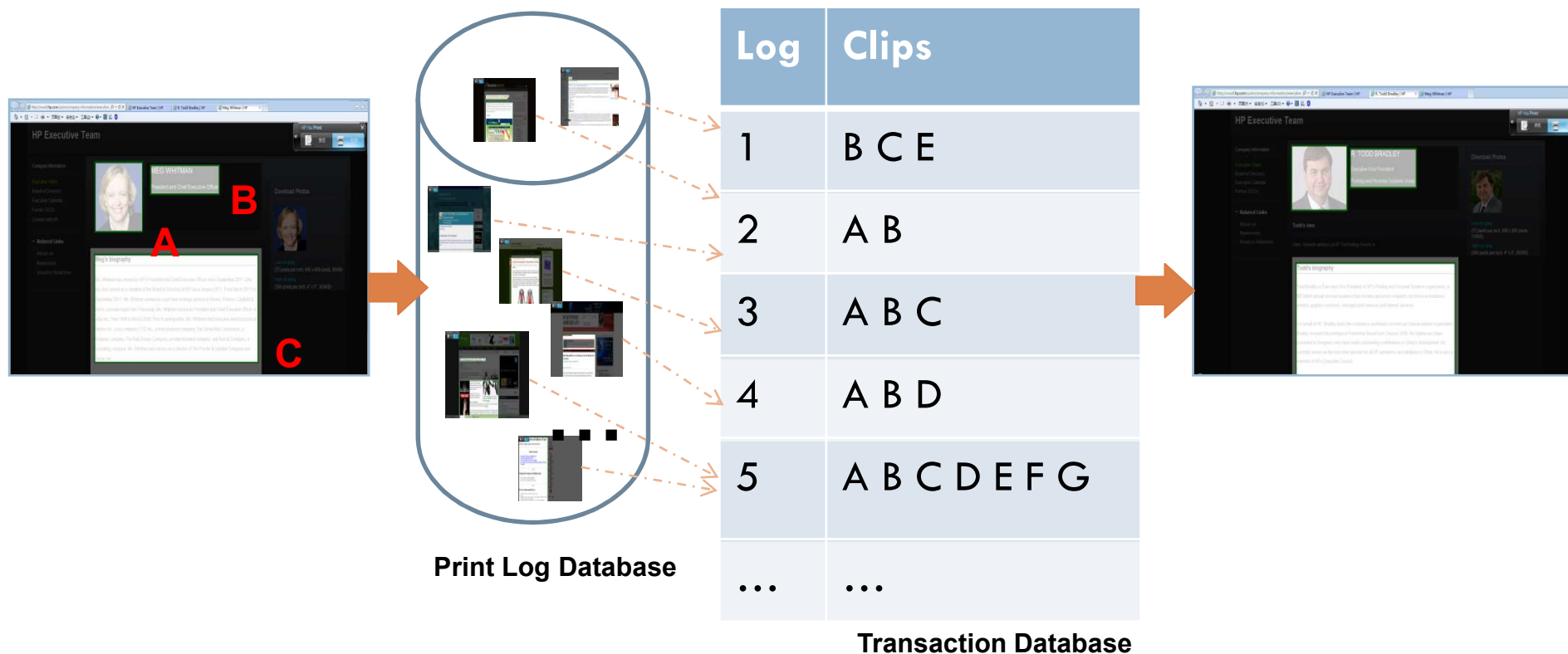




# 兴趣建模

23

- 非个性化兴趣模型
  - 无法识别单个用户的ID

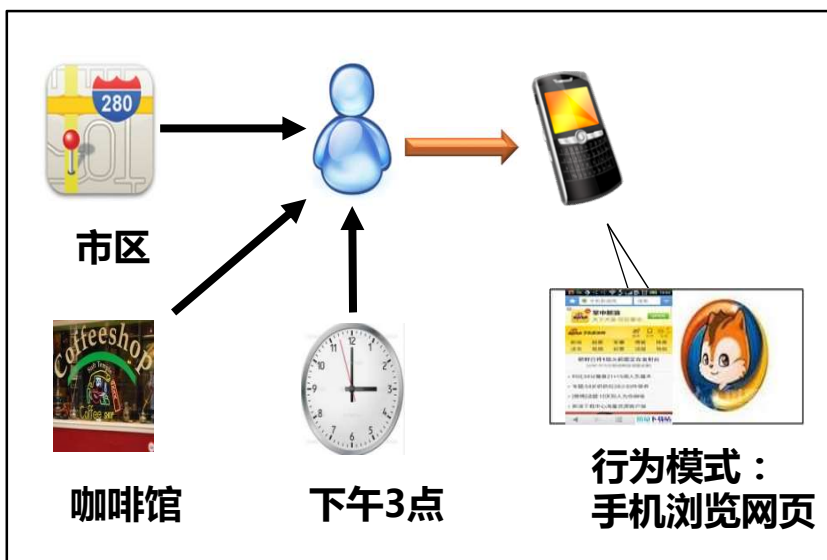




# 兴趣建模

24

- 非个性化兴趣建模
- 关联分析 ----发现经常出现的事物、行为、现象
  - 购买A的顾客还会购买。。。
  - 搜索“讯飞”的用户还会经常搜索“科大讯飞”
  - 移动用户在下午五点经常听歌还是玩游戏？
  - 下载某类音乐的用户通常有哪些特点？



用户行为与所处环境的关联

{(Is a holiday?:Yes),  
(time range: PM3:00-4:00), },  
(location: Urban)}

⇒ Surfing on the Internet



# 兴趣建模

25

## □ 频繁序列模式挖掘



SID	Search Session
1	丰田→雷诺→宝马→奔驰
2	宝马→奔驰→法拉利
3	本田→丰田→通用→宝马
4	吉利→奇瑞→长城→江淮
5	比亚迪→吉利→江淮→长安
6	长城→江淮→华泰→长安

搜索序列模式1:



意图:  
国际汽  
车品牌

搜索序列模式2:



意图:  
国内汽  
车品牌

## □ 类似于关联模式挖掘的算法

- PrefixSpan

## □ 隐马尔科夫模型、CRF

- 将搜索意图理解为隐状态

- 建模搜索历史与当前意图之间的序列模式



# 兴趣建模

26

- 个性化兴趣模型
  - 利用项目列表来表示用户兴趣
    - 加权 or 不加权
    - 数据异常稀疏

	Item-1	Item-2	Item-3	Item-4	Item-5	Item-6
User-1	4	*	2	5	*	*
User-2	3	2	1	*	*	3
User-3	*	2	*	3	*	4
User-4	*	3	3	5	4	*
User-5	5	*	3	4	*	*





# 兴趣建模

27

## □ 个性化兴趣模型

	Item-1	Item-2	Item-3	Item-4	Item-5	Item-6
User-1	4	*	2	5	*	*
User-2	3	2	1	*	*	3
User-3	*	2	*	3	*	4
User-4	*	3	3	5	4	*
User-5	5	*	3	4	*	*

Three kinds of data: users' data, items' data, and ratings' data.

- Name
- Job
- Like
- Address
- .....

- Name
- Tag
- [content]
- [category]
- [time]
- .....

- Value
- User-id
- Item-id
- [time]
- .....



# 兴趣建模

28

## 个性化兴趣模型

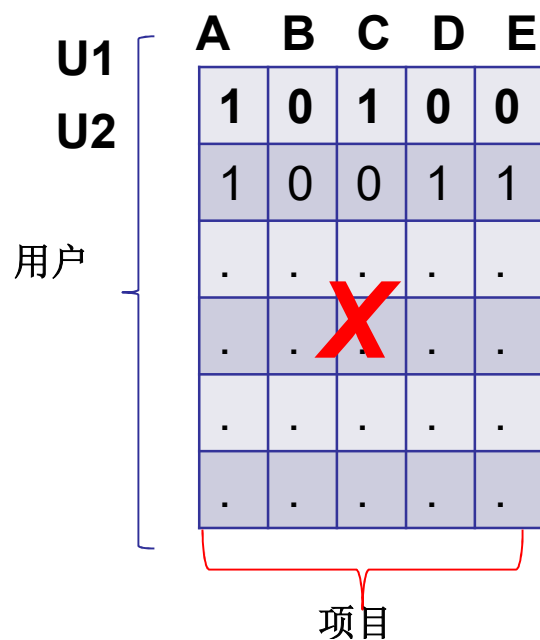
### 低秩分解

#### 评分预测

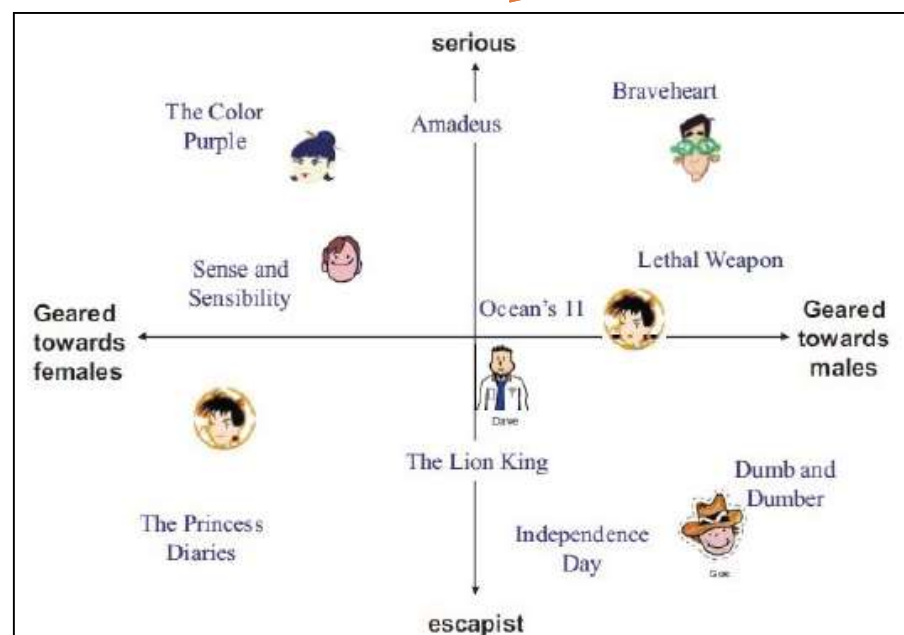
■ 1-5分

#### 学习排序

■ 1购买, 0未购买



	Item-1	Item-2	Item-3	Item-4	Item-5	Item-6
User-1	4	*	2	5	*	*
User-2	3	2	1	*	*	3
User-3	*	2	*	3	*	4
User-4	*	3	3	5	4	*
User-5	5	*	3	4	*	*





# 兴趣建模

29

## □ 显式信息(Explicit Information)

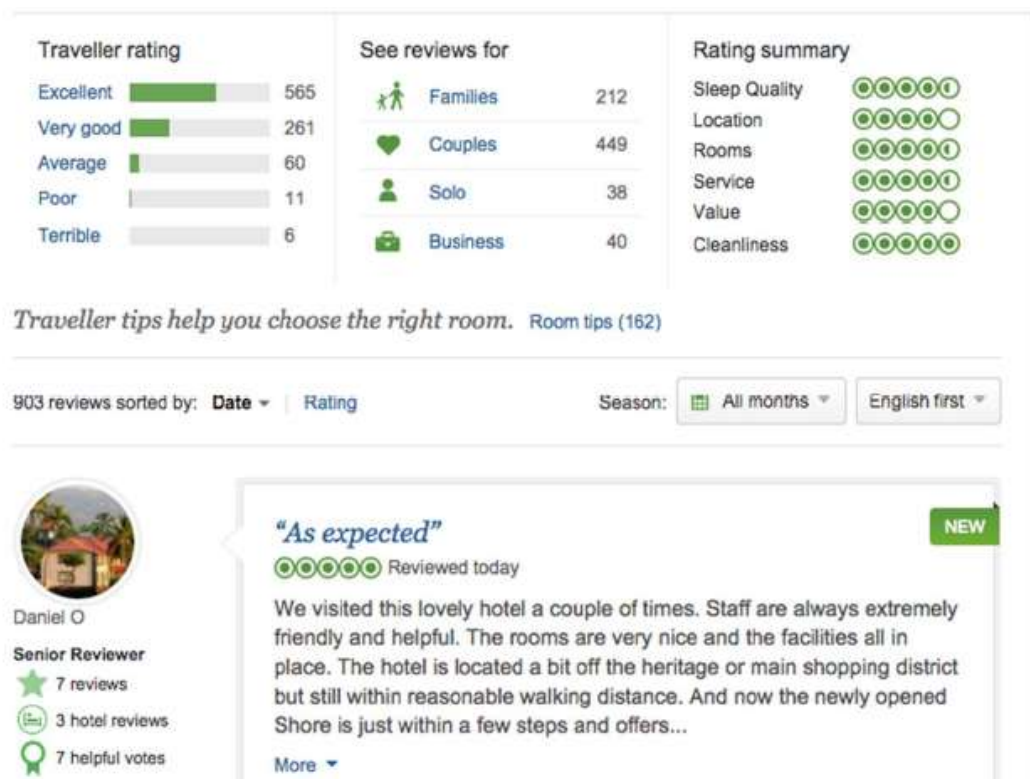
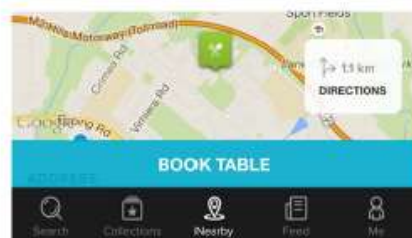
□ 用户手动添加，系统直接从用户输入获得

■ 评分、点赞、评论等



TABLE RESERVATION

Recommended





# 兴趣建模

30

## □ 隐式信息(Implicit Information)

□ 自动生成，要么是和系统的交互信息，要么是模型训练所得

■ 例如点击、浏览等日志行为

■ 还如鼠标轨迹、浏览器缓存等等



Table 1: A toy example of the customer behaviorial records.

Userld	Itemld	Categoryld	Action	Timestamp
$U_1$	a	$C_1$	Click	2014-07-08 20:05:20
$U_1$	b	$C_1$	Click	2014-07-08 20:06:40
...	...	...	...	...
$U_1$	a	$C_1$	Cart	2014-07-08 20:13:55
$U_1$	b	$C_1$	Collect	2014-07-08 20:14:20
$U_1$	b	$C_1$	Buy	2014-07-08 20:14:38
$U_2$	f	$C_2$	Click	2014-07-09 10:21:13
$U_2$	f	$C_2$	Buy	2014-07-09 10:21:20

Figure 1: An example automatically discovered motif from mouse



# 兴趣建模

31

## 融合情境信息的个性化兴趣模型

### 基于情境信息的用户行为切分

#### 移动情境



- 当前时间
- 用户位置
- 移动速度
- 周围噪声

。 。 。





# 兴趣建模

32

## 融合情境信息的个性化兴趣模型

### 基于情境信息的用户行为切分

#### ■ 旅游情境

#### 季节

#### 地区



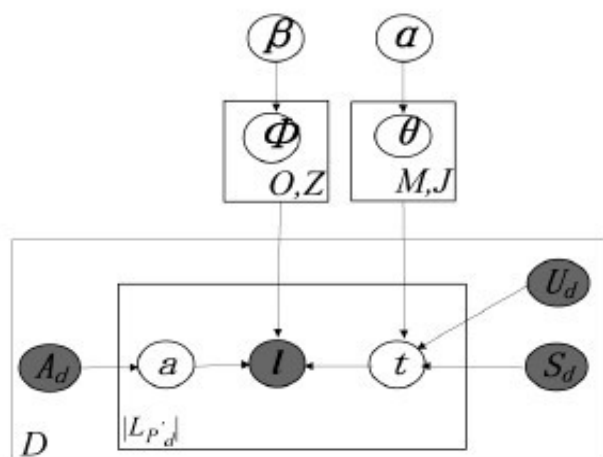




# 兴趣建模

33

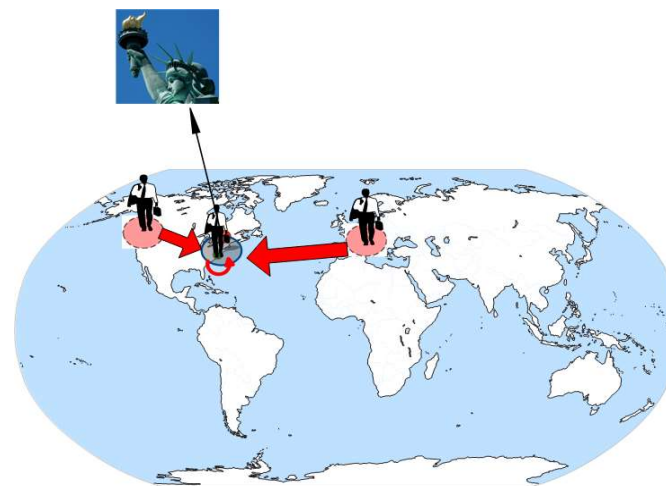
- 融合情境信息的个性化兴趣模型
  - 基于情境信息的用户行为切分
    - 旅游情境



情境感知的用户兴趣挖掘模型

$U_1$	$U_2$
0.27	0.05
0.10	0.02
0.05	0.27
0.27	0.02
0.02	0.27
...	...

用户兴趣表示





# 兴趣建模

34

## 融合情境信息的个性化兴趣模型





# 兴趣建模

35

## 融合情境信息的个性化兴趣模型





# 兴趣建模

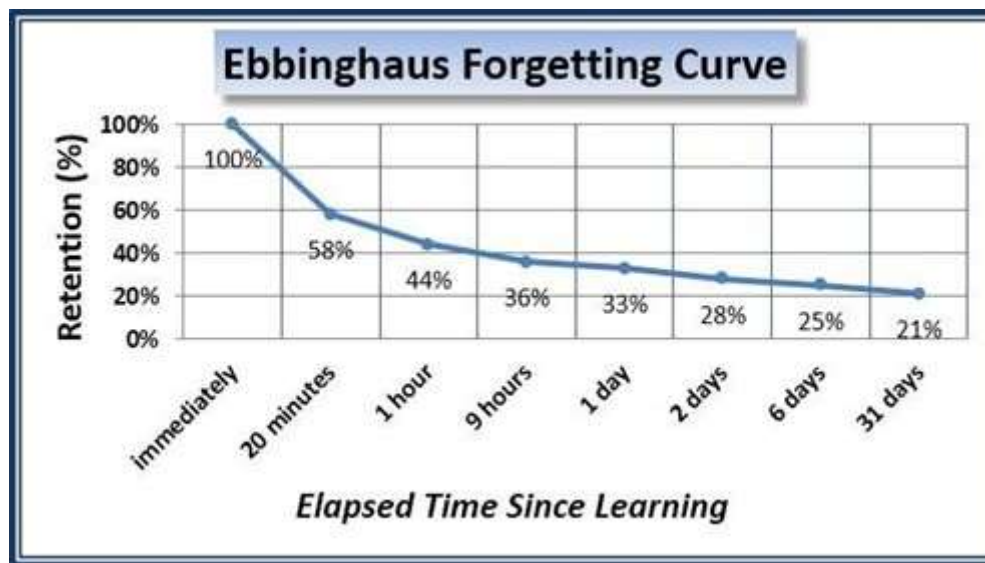
36

## 融合能力的个性化兴趣模型

### 基于能力的用户画像

#### 学习场景

- 知识点：数学、物理、词汇
- 技能：记忆力、逻辑、演算





# 兴趣建模

37

## 融合能力的个性化兴趣模型

### 基于能力的用户画像

#### 竞技场景

- 竞技技能：速度、力量、防守
- 竞技意识：协防、突袭、反击





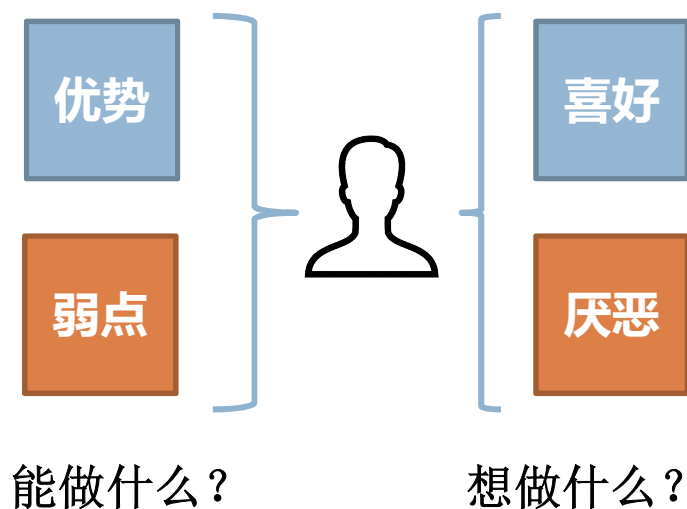
# 兴趣建模

38

## 融合能力的个性化兴趣模型

### 基于能力的用户画像

#### 能力 vs 兴趣



#### 匹配推荐

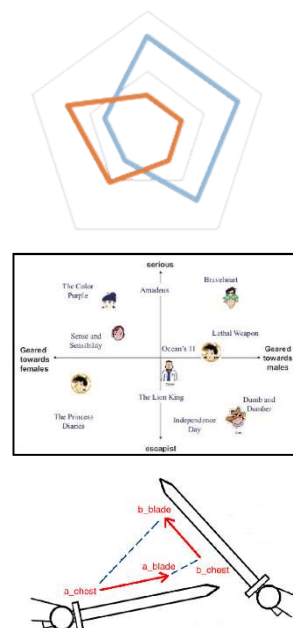
用户能力建模



用户兴趣建模



匹配推荐



匹配合适的项目或者合适的对手

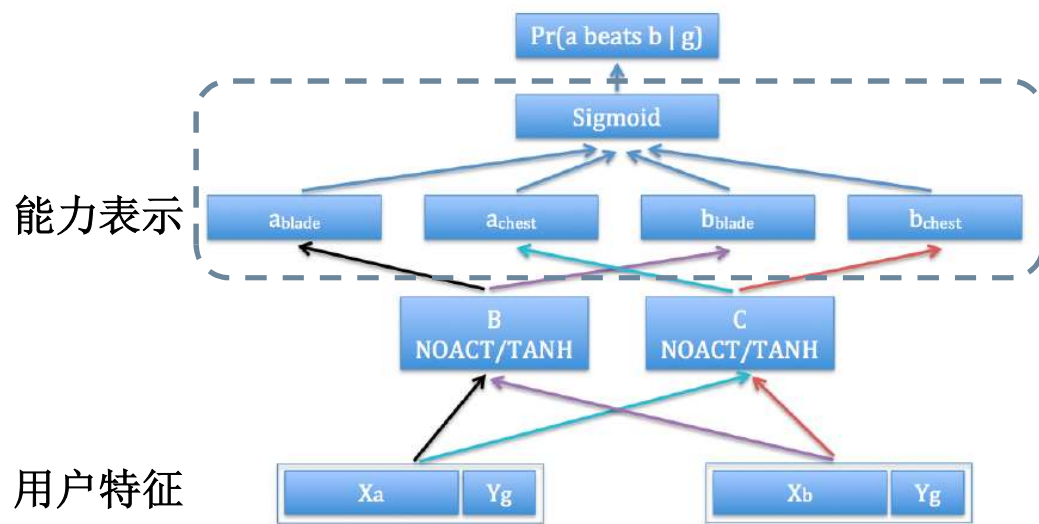




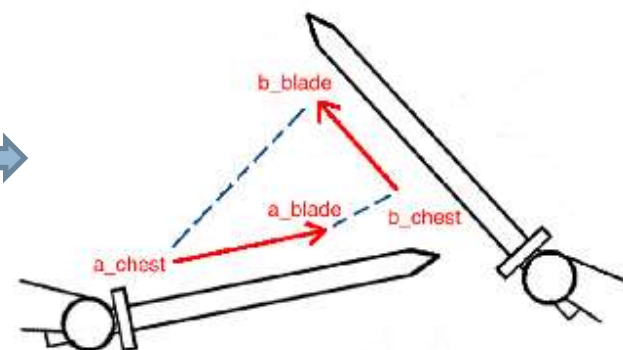
# 兴趣建模

39

## 融合能力的个性化兴趣模型



基于对抗的用户能力表示模型



Blade: 优势or锋芒  
Chest: 弱点or软肋

“锋芒-软肋” 对抗模型

$$\Pr(a \text{ beats } b) \propto ||\mathbf{b}_{\text{blade}} - \mathbf{a}_{\text{chest}}||_2^2 - ||\mathbf{a}_{\text{blade}} - \mathbf{b}_{\text{chest}}||_2^2$$