



中國人民大學
RENMIN UNIVERSITY OF CHINA

计算传播理论与实务

2019-2020秋季学期

第五讲

算法推荐系统简介

授课教师：范举副教授、塔娜讲师
时间：2019年12月2日



算法推荐

- ✓ 算法推荐系统概览
- ✓ 用户建模和分析
- ✓ 内容建模和分析
- ✓ 推荐算法

第5.1节

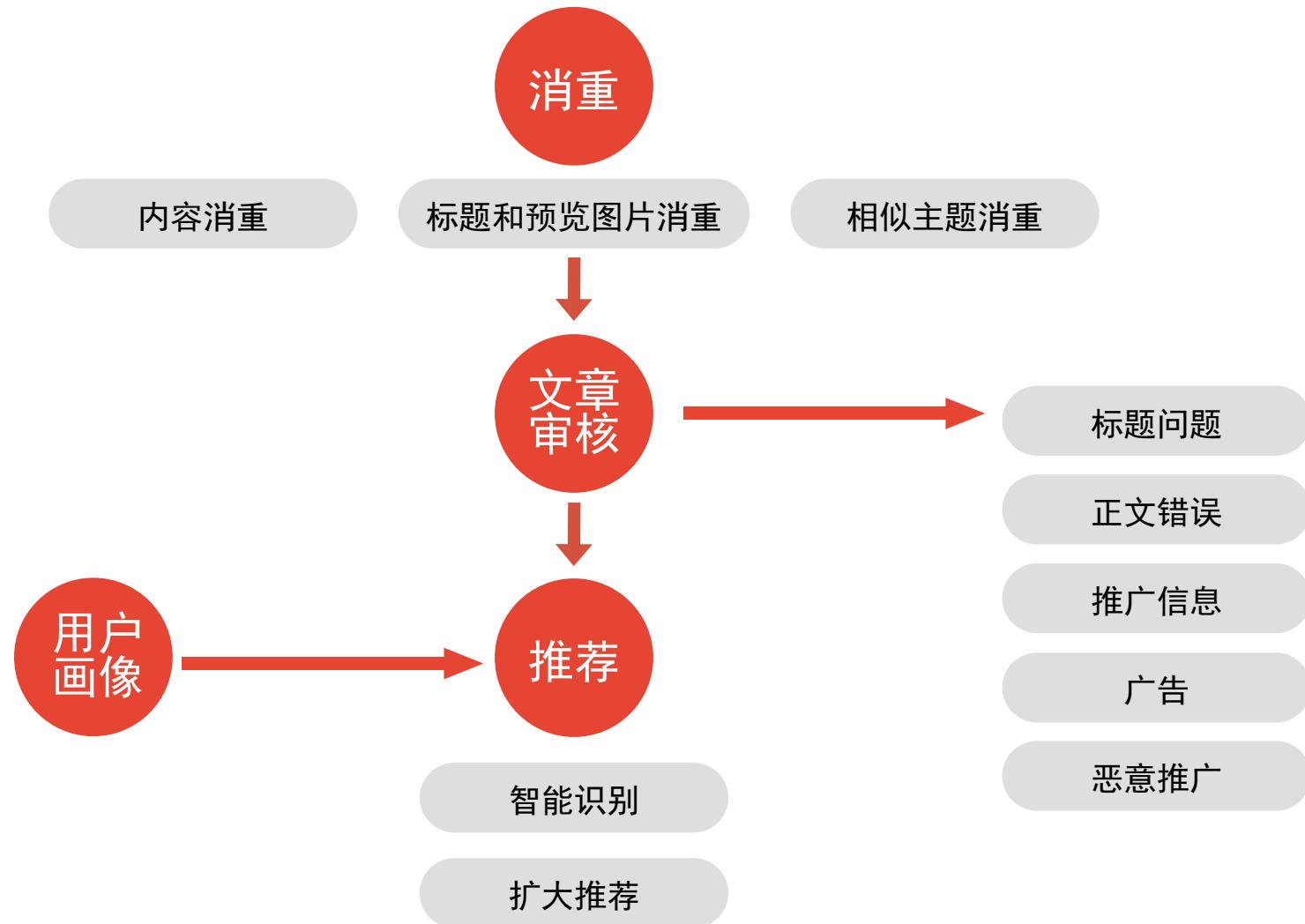
算法推荐系统概览

简化的算法分发模型



核心元素：用户、内容、分发逻辑

算法推荐的基本流程



核心问题

- 推荐系统本质上要解决 用户，环境和内容的匹配：

$$y = f(x_{\text{user}}, x_{\text{env}}, x_{\text{content}})$$

- 推荐
- 不推荐



引自：曹欢欢博士《今日头条推荐系统简介》

第5.2节

用户建模和分析



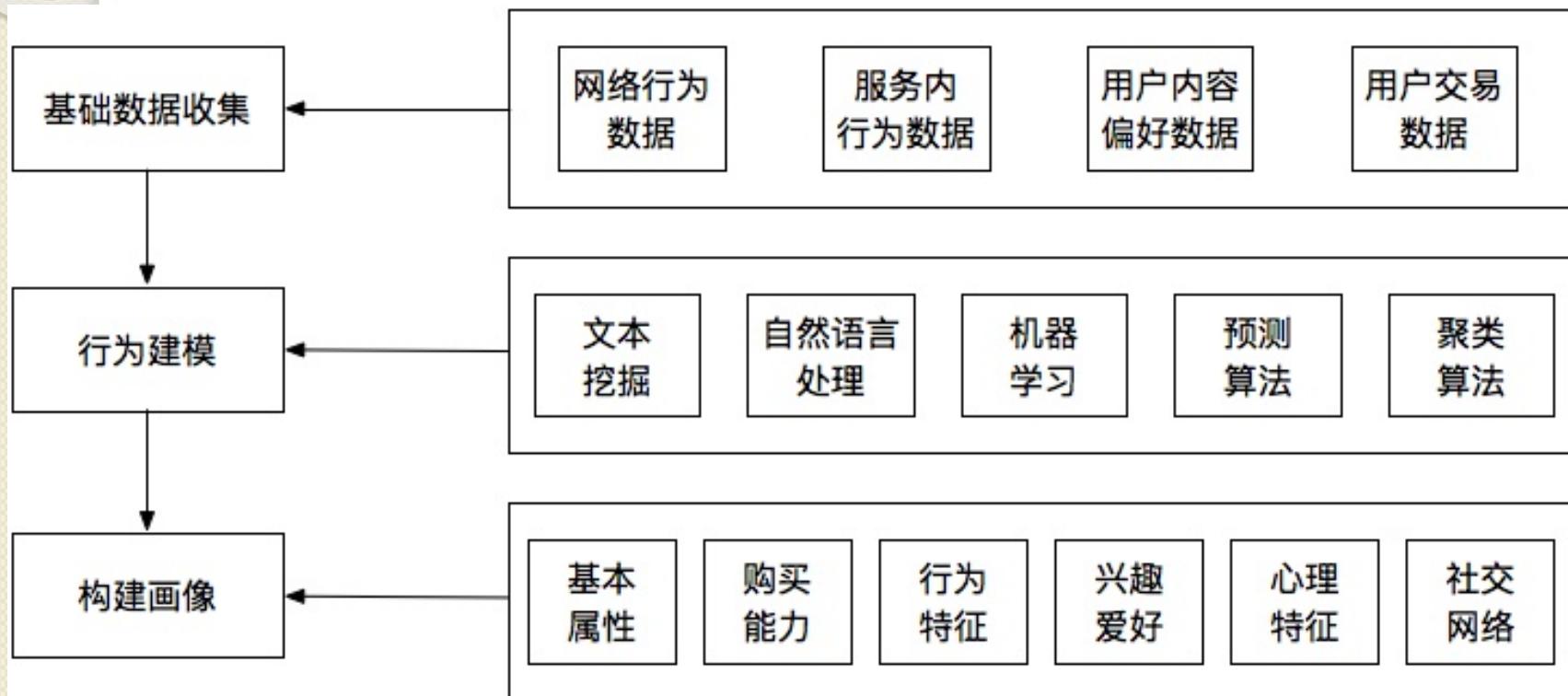
用户画像 (User profile/portrait)

- 根据用户的社会人口属性、生活习惯、消费行为等信息而抽象出的一个标签化用户模型
- 构建用户画像的核心工作是给用户打“标签”
- “标签”是对用户信息分析得来的高度精炼的特征标识
- 举例：经常购买玩具→“有孩子”→“有5-8岁孩子”

用户画像的作用

- 精准营销：分析产品潜在用户，定向特定群体
- 用户统计：中国大学购买书籍人数TOP10
- 数据挖掘，智能推荐：利用关联规则计算，喜欢红酒的人通常喜欢什么运动品牌
- 效果评估：完善产品运营，提升服务质量
- 服务/产品的私人定制：个性化服务某类群体甚至每一位用户

用户画像的构建流程



用户标签体系举例

模型：

- 每个用户就是一组标签的集合
- 标签可以设置不同的权重

兴趣特征

- 感兴趣的类别和主题
- 感兴趣的关键词
- 感兴趣的来源
- 基于兴趣的用户聚类
- 各种垂直兴趣特征（车型，体育球队，感兴趣股票）

身份特征

- 性别
- 年龄
- 常驻地点

行为特征

- 晚上才看视频



用户画像的初始化——怎样解决“冷启动”？

- 新用户：标签数据少，推荐效果不好
- 扩充数据来源：通过微博/微信登陆，获取更多信息
- 手机的机型，用户的位置，……

为用户设置/调整标签有哪些策略

- 过滤噪声：过滤停留时间短的点击，打击标题党
- 惩罚热点：用户在热门文章上的动作做降权处理
- 时间衰减：随着用户动作的增加，老的特征权重会随着时间衰减，新动作贡献的特征权重会更大
- 惩罚展现：如果一篇推荐给用户的没有被点击，相关特征（类别，关键词，来源）权重会被惩罚
- 考虑全局背景：考虑给定特征的人均点击比例

第5.3节

内容建模和分析

文本分析

- 对文本的表示及其特征项的选取
- 文本挖掘、信息检索的一个基本问题，使用从文本中抽取出的特征词进行量化，表示文本信息
- 非结构化 → 结构化，可处理的
- 基本技术：分词、词频统计
- 向量空间模型：描述文本向量
- 向量空间降维：特征（feature）选择
- 例：我/在/中国人民大学/读博士。我/在/上海/开会。

文本分析在推荐系统中的作用

- 用户兴趣建模 (user profile) : 如, 给喜欢阅读【互联网】文章的用户打上【互联网】标签, 给喜欢【小米】新闻的用户打上【小米】标签
- 优化内容组织: 如生成频道内容, 【德甲】的内容进【德甲频道】, 【瘦身】的内容进【瘦身频道】
- 帮助内容推荐: 【魅族】的内容推荐给关心【魅族】的用户, 【Dota】的内容推荐给关心【Dota】的用户

文本特征举例：类别，关键词

查找文章:

[4688699423 莎娃连续17次不敌小威 07-10 13:18 rate:18 展开>>](#)

.....

文章Profile

一级分类 展开>>	
news_sports	2.5957

二级分类 展开>>	
news_sports/tennis	0.7201

关键词2 展开>>							
西班牙	0.9915	小威	0.9858	穆古拉扎	0.9845	女单决赛	0.9641
俄罗斯	0.9475	莎拉波娃	0.9282	莎娃	0.9208	小威廉姆斯	0.9199
委内瑞拉	0.8738	锦标赛	0.7582	温网	0.6409	大满贯	0.5660
半决赛	0.4663						

高亮关键词 展开>>							
西班牙	0.9976	莎拉波娃	0.9886	俄罗斯	0.9856	小威廉姆斯	0.9831
委内瑞拉	0.9823	小威	0.9498	穆古拉扎	0.9463	温网	0.9323
半决赛	0.7198	女单决赛	0.7114	大满贯	0.6948	波兰	0.6094

文本特征举例：话题/topic

查找文章: 4688699423

[4688699423 莎娃连续17次不敌小威 07-10 13:18 rate:18 展开>>](#)

.....
.....

文章Profile

一级分类 展开>>	
news_sports	2.5957

二级分类 展开>>	
news_sports/tennis	0.7201

关键词2 展开>>							
西班牙	0.9915	小威	0.9858	穆古拉扎	0.9845	女单决赛	0.9641
俄罗斯	0.9475	莎拉波娃	0.9282	莎娃	0.9208	小威廉姆斯	0.9199
委内瑞拉	0.8738	锦标赛	0.7582	温网	0.6409	大满贯	0.5660
半决赛	0.4663						

高亮关键词 展开>>							
西班牙	0.9976	莎拉波娃	0.9886	俄罗斯	0.9856	小威廉姆斯	0.9831
委内瑞拉	0.9823	小威	0.9498	穆古拉扎	0.9463	温网	0.9323
半决赛	0.7198	女单决赛	0.7114	大满贯	0.6948	波兰	0.6094

文本特征在推荐过程中的特殊作用

- 没有文本特征， 推荐引擎无法工作
- 粒度越细的文本特征， 冷启动能力越强， 如
【拜仁慕尼黑】 VS 【体育】

文本分析算法举例：实体词识别 算法

英超-利物浦0-0曼联，德赫亚频频开挂

原创 肆客足球 2016-10-18 07:54

北京时间10月18日凌晨03:00，2016-17赛季英超联赛第八轮焦点战打响，红军利物浦坐镇安菲尔德球场迎战红魔曼联，上演两队第197次双红会。上半场，红军采用高压反抢限制曼联进攻，在高空球方面，曼联则占据优势。半场双方互无建树。易边再战，双方攻势渐起，德赫亚两次神扑将利物浦极具威胁的进攻化解。全场战罢，双方0-0握手言和。积分榜上，利物浦落后榜首的曼城2分排在第4，曼联积14分排在第7位。

分词&词性标注

英超 N 利物浦 N 0-0 曼联 N , 德赫亚 N

.....

抽取候选

新版实体词	展开>>
大卫·德赫亚	0.9973
利物浦足球俱乐部	0.9899
曼彻斯特联足球俱乐部	0.9835
英格兰足球超级联赛	0.9565
兹拉坦·伊布拉希莫维奇	0.6718
卢克·肖	0.6559
韦恩·鲁尼	0.6387
埃姆雷·詹	0.6320
保罗·博格巴	0.6196
迈克尔·卡里克	0.5185

计算相关性

英超联赛
利物浦足球俱乐部*
利物浦市*
曼联俱乐部
德赫亚
.....

去歧

英超联赛
利物浦足球俱乐部
曼联俱乐部
德赫亚
.....

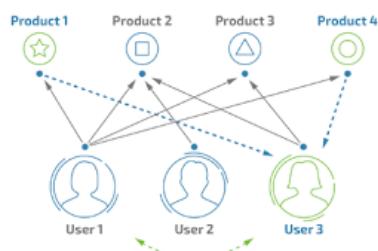
第5.4节

推荐算法

讨论

- 你常用的算法推荐系统中，是依赖用户自身数据推荐，还是引入了社交关系，你觉得那个占比更大？
 - 微博，朋友圈
 - 淘宝，小红书
 - 抖音，快手

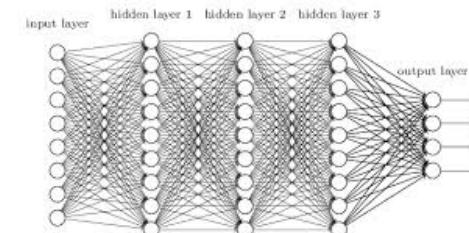
典型的推荐算法



协同过滤

$$F(x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$$

逻辑回归

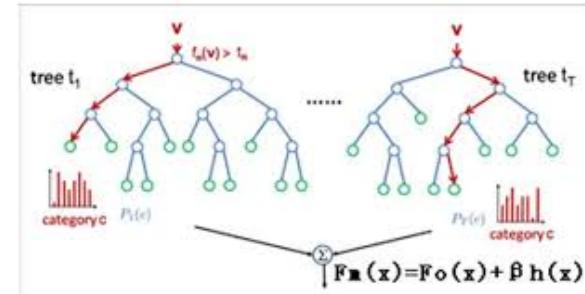


深度神经网DNN

$$\hat{y}(x) = w_0 + \sum_{j=1}^p w_j x_j + \sum_{j=1}^p \sum_{j'=j+1}^p x_j x_{j'} \langle v_j v_{j'} \rangle$$

$$\hat{y}(x) = w_0 + \sum_{j=1}^p w_j x_j + \sum_{j=1}^p \sum_{j'=j+1}^p x_j x_{j'} \sum_{f=1}^k v_{f,j} v_{f,j'}$$

因子分解机
Factorization
Machine



梯度提升树GBDT

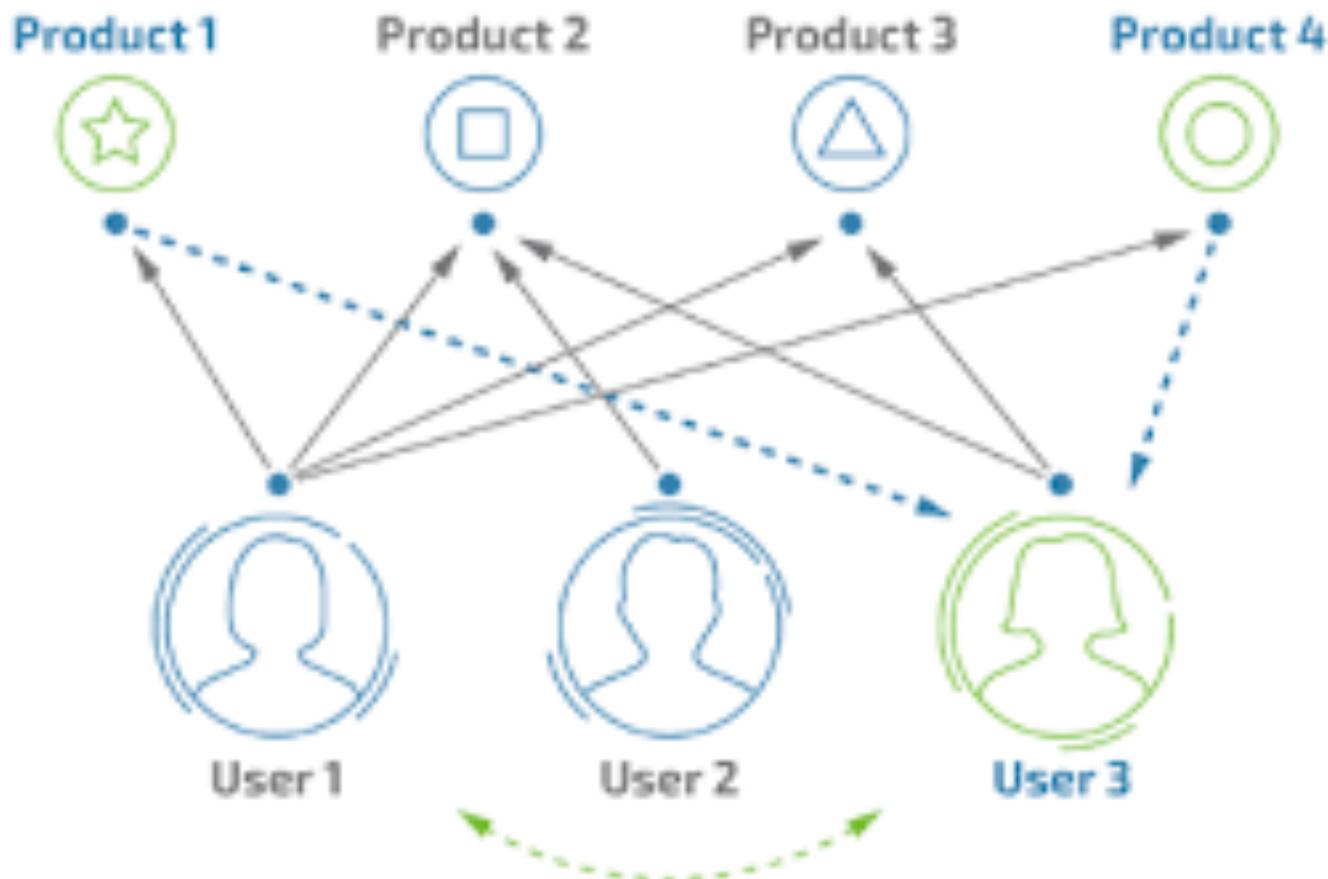
协同过滤算法 (Collaborative Filtering)

- 推荐系统领域最基本、应用最为广泛的算法
- 通过分析和利用用户的历史行为来给用户的兴趣建模，并根据用户的兴趣为用户做出推荐。
- 基于用户CF：最早被应用于邮件过滤和新闻推荐中
- 基于物品CF：最早由Amazon推荐系统的专家提出，在商业界广泛应用

用户CF

- 基本假设：一个用户会喜欢和他有相似兴趣、喜好的用户群喜欢的物品
- 为了给目标用户做推荐，首先应该找到与该用户在兴趣喜好上最相似的一组用户，然后做推荐
- 两个用户相似是指这两个用户喜欢过的物品集合相似

用户CF模型



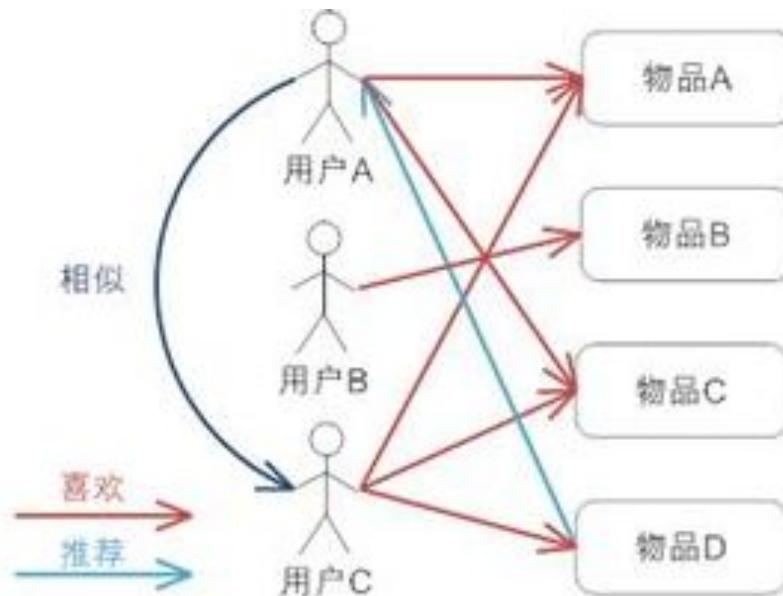
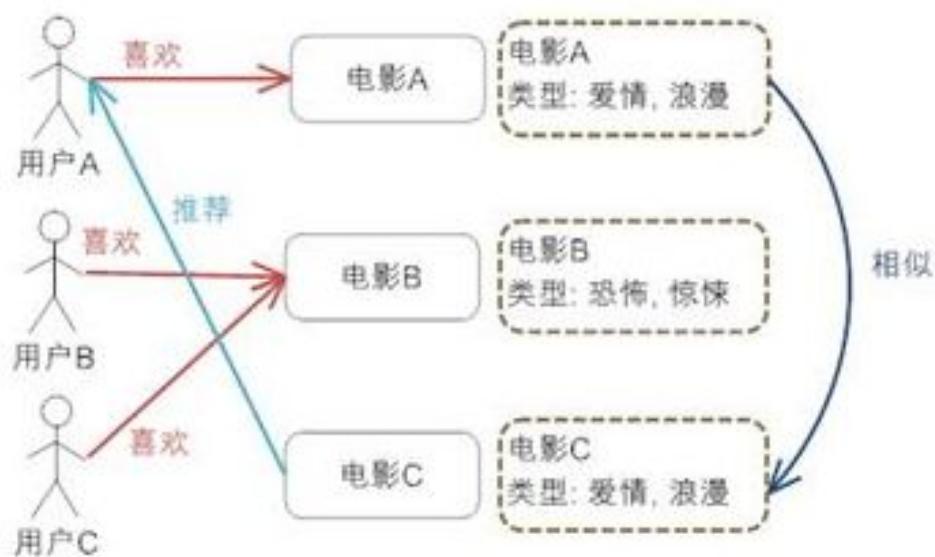
基于物品的协同过滤算法

- 基本假设：用户会喜欢和他以前所喜欢的物品相似的物品
- 推荐时，首先从用户行为历史数据中检索他之前喜欢过的物品集合，然后从尚未推荐的物品里找到和他喜欢过的物品相似的物品，进行推荐
- 两个物品相似是指喜欢过这两个物品的用户集合相似

比较

- 基于用户：更加社会化，反映基于社交关系的兴趣
- 新闻推荐
- 基于物品：更加个性化，反映用户自身喜好
- 电商推荐
- 融合使用

比较



试一试

如下数据是各用户对各文档的偏好：

用户/文档	文档A	文档B	文档C	文档D
用户A	√	√	推荐？	推荐？
用户B	√	√		√
用户C	√		√	√
用户D	√			√

现在需要基于上述数据，给A用户推荐一篇文档

推荐系统的数据依赖

- 推荐模型的特征抽取需要用户和内容的各种标签
- 召回策略需要获取用户侧和内容侧的各种标签
- 用户标签挖掘和内容分析是搭建推荐系统的基石

课程总结

- 算法推荐系统
- 用户建模
- 内容建模
- 推荐算法



Thanks!