



大数据成就未来



智能推荐


姜鹏辉

目录

1	电子商务网站智能推荐服务
2	智能推荐概述
3	协同过滤算法
4	上机实验
5	拓展思考



商品已成功加入购物车！



R语言实战 第2版
颜色：R语言实战 第2版 / 数量：1



举报

55个卖家在售 **¥64.20** 起



案例背景

全部订单

待付款

待收货

待评价3

我的常购商品NEW


近三个月订单

订单详情

2016-10-19 00:13:49

订单号: 39979501033


京东



Python核心编程 (第3版)

x1

找搭配



金字塔原理大全集 (套装共2册)


x1

找搭配

2016-09-10 18:02:42

订单号: 23122141663

河北



小米 (MI) 5000Ah移动电源超薄锂聚合物手机通用充电宝 银色 5000电源+橙色

x1

找搭配

2016-08-02 20:40:18

订单号: 22561711070

吉吉

商品评价

99%

好评度

好评(99%)

中评(1%)

差评(0%)

全部评价(400+)

晒图(96)

好评(400+)

中评(4)

差评(1)

只看当前商品评价


★★★★★


学习R的入门基础书, 不错的选择!


购买21天后评价


2016-08-24 12:47

R语言实战 第2版










点赞(0)

回复(0)

★★★★★

通过这个R语言, 学习数据分析, 满满看


猜你喜欢



用Python写网络爬虫

¥37.70


(已有17652人评价)



利用Python进行数据分析

¥75.70


(已有2968人评价)



Python绝技: 运用Python成为顶级黑客

¥66.20

(已有2820人评价)



影响力 (经典)

¥35.50

(已有41656人评价)

案例背景

- 常见推荐方式
- 热点推荐
- 经常一起购买的产品：打包销售
- 购买此产品的顾客同时也购买了：协同过滤 - 显式需求
- 看过此商品后顾客购买的其他商品：协同过滤 - 隐式需求
- 用户评论（打分）列表
- 亚马逊20%~30%销售额来自推荐系统
- 几乎所有大型电子商务网站10%以上销售额来自推荐系统



案例背景

用户行为	类型	特征	作用
评分	显式	整数量化的偏好，可能的取值是[0, n]	通过用户对物品的评分，可以精确得到用户的偏好
投票	显式	布尔量化的偏好，取值是0或1	通过用户对物品的投票，可以较精确地得到用户地偏好
转发	显式	布尔量化的偏好，取值是0或1	通过用户对物品的投票，可以精确得到用户的偏好。如果是站内，同时可以推理得到被转发人的偏好（不精确）
保存书签	显式	布尔量化的偏好，取值是0或1	通过用户对物品的投票，可以精确得到用户的偏好
标记书签(Tag)	显式	一些单词，需要对单词进行分析，得到偏好	通过分析用户的标签，可以得到用户对项目的理解，同时可以分析出用户的情感：喜欢还是讨厌
评论	显式	一段文字，需要进行文本分析，得到偏好	通过分析用户的评论，可以得到用户的情感：喜欢还是讨厌
单击流（查看）	隐式	一组用户的点击，用户对物品感兴趣，需要进行分析，得到偏好	用户的单击一定程度上反映了用户的注意力，所以它也可以从一定程度上反应用户的偏好
页面停留时间	隐式	一组时间信息，噪声大，需要去噪，分析得到偏好	用户的页面停留时间一定程度上反映了用户的注意力和偏好，但噪声偏大，不好利用
购买	显式	布尔量化的偏好，取值是0或1	用户的购买行为很明确地说明他对这个项目感兴趣

案例背景

某法律网站是一家大型的法律资讯信息网站，它一直致力于为用户提供丰富的法律资讯信息与专业法律咨询服务，并为律师与律师事务所提供卓有成效的互联网整合营销解决方案。



中国法律资讯门户
[城市切换]全国

咨询 马上搜索 免费咨询

地区站点：广东站 广州站 深圳站 珠海站 汕头站 韶关站 河源站 梅州站 惠州站 汕尾站 东莞站 中山站 江门站 佛山站 阳江站 湛江站 茂名站 肇庆站 云浮

找律师 按地区 按专业 按律所 法律咨询 HOT 咨询列表 咨询标签 提交咨询 法律知识 婚姻知识 劳动知识 房产知识 法律法规 民事法规 刑事法规 经济法规 公检法 立法动态 法院资讯 司法行政 案件委托 发布委托 法律协助 委托信息 法律资讯 法制要闻 法律生活 法律时评 律师聚焦 律界热点 新案关注 立法草案

专业频道 婚姻法 刑法 劳动法 房地产法 合同法 公司法 交通事故 破产法 合同范本 保险法 著作权法 商标法 专利法 继承法 民事诉讼 刑事诉讼 更多

向全国十万律师提问
请在此输入您的咨询内容，祝您问题早日得到解决!
解答咨询量：1039855条 提交咨询
12分钟前 李佳律师 回复了法律咨询
12分钟前 喻远军律师 回复了法律咨询

广州 肖艳平律师
TEL：185-2064-8558

广州 练武律师
法律咨询：13416225037

广州 梁火榮律师
法律咨询：13480258386

中国最大的法律资讯与律师门户
the award-winning website of law

许胜利律师 TEL：135-3380-6942
承办过大量各类典型疑难案件，积累了丰富的办案经验。

 泰迪智能科技
TipDM Intelligent Technology

大数据挖掘专家

7

案例背景

目前网站上已经存在部分推荐，比如：当访问主页时可以在婚姻栏目发现如下热点推荐。

婚姻法热文

- 01 协议离婚后 反悔 12-27
- 02 离婚后财产纠纷案例及依据 12-27
- 03 离婚时分割房产的几个问题 12-27
- 04 明智女人选择婚前协议 12-27
- 05 我想离婚，但不知道怎么办， 12-27
- 06 分居两年能离婚吗 12-27
- 07 结婚的特征是什么？ 12-27
- 08 罪犯在被管制或缓刑期间能否结 12-27
- 09 婚姻关系存续期间，夫妻一方以 12-27
- 10 事实婚姻还是非法同居 11-30

婚姻法律咨询

我想离婚、要到儿子的抚养权

- 家庭暴力
- 彩礼
- 男方有外遇，提出离婚，财产怎么分割？小孩判给谁？

婚姻法律知识

- 重婚罪 | 重婚罪的犯罪嫌疑人追究刑事责任的程序有两种
- 抚养费 | 抚养费的支付标准和支付方式
- 彩礼 | 彩礼应否返还受关注 应返还的情形
- 感情破裂 | 离婚时如何认定系夫妻感情破裂

当访问具体的知识页面时，可以在页面的右边以及下面发现也存在一些热点推荐和基于内容的关键字推荐

协议离婚后 反悔

作者: JONER 来源: 未知 2013-12-27 18:24

离婚双方在婚姻登记处办理登记手续后，一方又反悔的，能否向法院起诉？这种情形下，一般可作如下几种判断：

- 1、一方又不同意离婚的，法院不予受理。双方要恢复婚姻关系的，重新向婚姻登记处申请复婚登记。
- 2、一方对离婚协议中的内容反悔的，可以在一年以内向人民法院起诉要求撤销，法院在查清签订离婚协议时没有欺诈，胁迫等情形的，裁定驳回起诉。
- 3、一方对子女的要求变更的可以向人民法院起诉。如不具有变更权的正当理由，法院驳回诉讼请求。
- 4、要求增加抚养费的，可以向法院起诉。

相关知识推荐

- 在法国提出离婚会对居留产生不利影响
- 夫妻双方离婚时保险财产如何分割
- 离婚案件中缺席判决的适用
- 公司股份是否应当作为婚前财产进行分割
- 配偶权与离婚精神损害之间的关系
- 离婚必须双方到场吗
- 外地人在北京怎么离婚

推荐阅读

离婚后孩子归谁抚养 有优先条件
离婚两年后，经济条件改善不能成为变更抚养权
判决离婚有哪些法定条件
关于离婚财产分割后逃债的问题【案例详解】
离婚后“夫妻”间给予经济帮助的条件
婚姻登记处办理离婚登记的条件是什么
离婚如何取证

学会主动用法律保护自己的权益
学习法律知识

热文其他离婚知识

- 1 社会抚养费征收程序
- 2 结婚证撕了怎么办
- 3 起诉离婚要提供哪些证据
- 4 非婚生子女户口



案例背景

目前情况：

- 目前网页上是基于内容的推荐（通过关键词）以及非个性化推荐。
- 推荐页面位置不明显，长篇的法律知识的末端。

婚姻法司法解释三发布（全文）

2014-11-05 | 作者：z

5096人

核心内容：最新婚姻法司法解释三于2011年7月4日由最高人民法院审判委员会第1525次会议通过，并于2011年8月13日起施行。下面法律快车婚姻法小编为您详细介绍婚姻法司法解释三的内容。

相关文章阅读

咨询专业律师

相关咨询推荐

- 婚姻法司法解释三全文 [来源：国家法规政策]
- 2011新婚姻法全文一婚姻法司法解释三 [来源：婚姻动态]
- 婚姻法司法解释三被误读 [来源：婚姻动态]
- 婚姻法司法解释三发布 多条款涉及房产确权 [来源：婚姻法规]
- 最高法发布19条婚姻法司法解释三 [来源：婚姻法规]
- 婚姻法司法解释三 2014 [来源：婚姻法规]

最新热门法律经验推荐

- | | | |
|----|----------------|--------|
| 1 | 2014驾驶证扣分新规定 | 227952 |
| 2 | 火车上能带酒吗？ | 190369 |
| 3 | 涨工资最新消息2014 | 187821 |
| 4 | 2014年最低工资标准是多 | 187573 |
| 5 | 2014最新劳动法产假规定多 | 164450 |
| 6 | 坐高铁可以带酒吗？ | 152436 |
| 7 | 2014年广东最低工资标准是 | 149284 |
| 8 | 2014天津最低工资标准是多 | 146537 |
| 9 | 2014上海最低工资标准 | 144765 |
| 10 | 江苏最低工资标准2014 | 135710 |



案例背景

面临以下问题：

- 1.访问用户多，是机会也是瓶颈；
- 2.留住用户，推荐律师；
- 3.自身推荐效果不佳。



挖掘目标

为了能够更好的满足用户需求，依据其网站海量的数据，研究用户的兴趣偏好，分析用户的需求和行为，发现用户的兴趣点，从而引导用户发现自己的信息需求。

挖掘目标：

- 按地域研究用户访问时间、访问内容、访问次数等分析主题，深入了解用户对访问网站的行为和目的及关心的内容。
- 借助大量的用户的访问记录，对不同需求的用户进行相关的服务页面的推荐。



原始数据

行为记录

realIP	realAreacode	userAgent	userOS	userID	clientID	timestamp
2683657840	140100	Mozilla/5.0 (Windows NT 5.1) AppleWebKit/537.36 (KHTML...	Windows XP	785022225.1422973265	785022225.1422973265	1422973268278
973705742	140100	Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKit/537...	Windows 7	2048326726.1422973286	2048326726.1422973286	1422973268308
3184681075	140100	Mozilla/5.0 (Windows NT 5.1) AppleWebKit/537.36 (KHTML...	Windows XP	1639801603.1422973278	1639801603.1422973278	1422973277375
3184681075	140106	Mozilla/4.0 (compatible; MSIE 8.0; Windows NT 5.1; Tride...	Windows XP	1597050740.1422973305	1597050740.1422973305	1422973282739
2683657840	140100	Mozilla/5.0 (Windows NT 5.1) AppleWebKit/537.36 (KHTML...	Windows XP	785022225.1422973265	785022225.1422973265	1422973290048
207452174	140100	Mozilla/5.0 (Windows NT 5.1) AppleWebKit/537.36 (KHTML...	Windows XP	589522884.1422272394	589522884.1422272394	1422973295258
432282638	140100	Mozilla/5.0 (Windows NT 6.1) AppleWebKit/537.36 (KHTML...	Windows 7	225597321.1422972988	225597321.1422972988	1422973305001
285097530	140100	Mozilla/5.0 (iPad; CPU OS 8_1_3 like Mac OS X) AppleWebKit/...	iOS	2105429197.1422973314	2105429197.1422973314	1422973309154
776247310	140100	Mozilla/5.0 (iPhone; CPU iPhone OS 8_1_1 like Mac OS X; ...	iOS	1577666249.1422457401	1577666249.1422457401	1422973317133
1275347569	140100	Mozilla/5.0 (Windows NT 6.1) AppleWebKit/537.36 (KHTML...	Windows 7	303317099.1422972785	303317099.1422972785	1422973319480
1768232564	140100	Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKit/537...	Windows 7	385670441.1422973098	385670441.1422973098	1422973321566
2891940471	140100	Mozilla/5.0 (Windows NT 5.1) AppleWebKit/537.36 (KHTML...	Windows XP	1468293794.1422972132	1468293794.1422972132	1422973324242
2962015864	140100	Mozilla/5.0 (Windows NT 6.1) AppleWebKit/537.36 (KHTML...	Windows 7	33209881.1417416558	33209881.1417416558	1422973328316
2977395726	140100	Mozilla/5.0 (Windows NT 6.1; WOW64; Trident/7.0; rv:1...	Windows 7	946774240.1422972832	946774240.1422972832	1422973330427
4207369840	140100	Mozilla/5.0 (Windows NT 6.1) AppleWebKit/537.36 (KHTML...	Windows 7	256955339.1422876538	256955339.1422876538	1422973333233

原始数据：

1、SQL文件

2、全国6个月数据

3、数据量250G

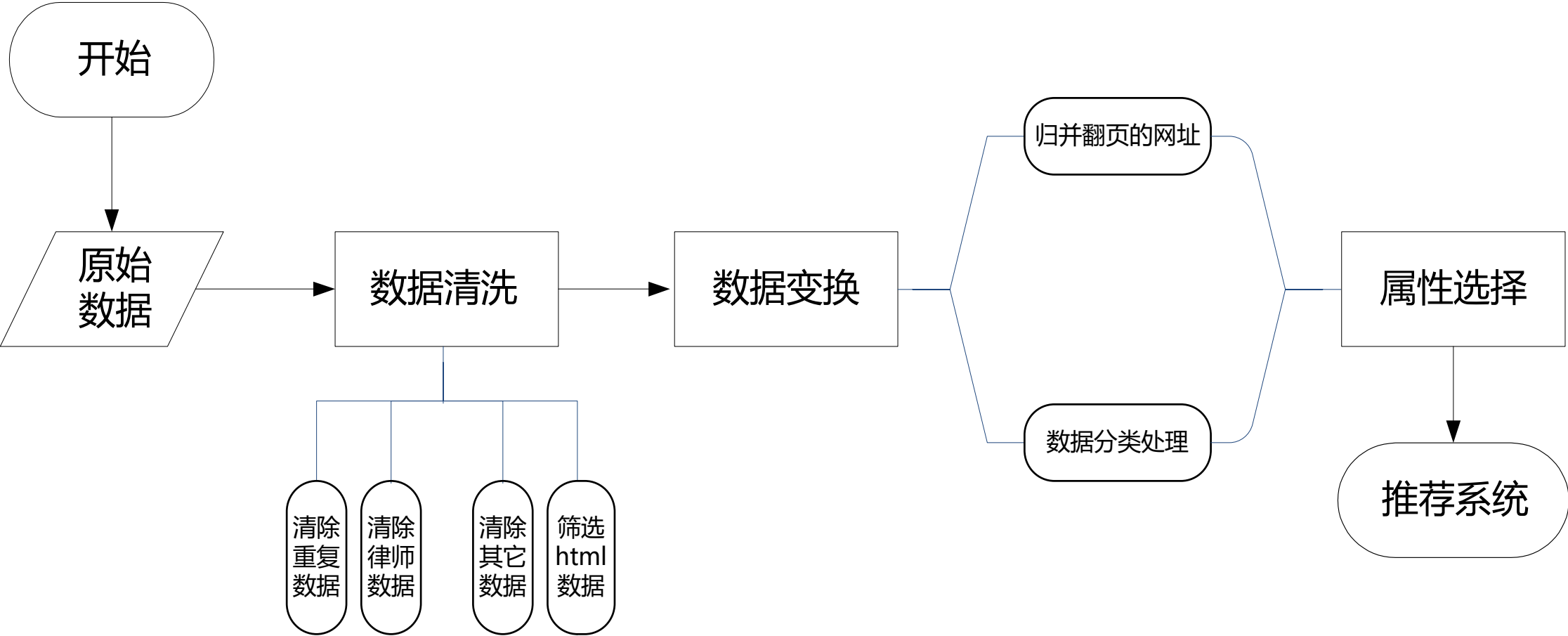
原始数据

当用户访问网站页面时，系统会记录用户访问网站的日志，其中记录了用户IP（已做数据脱敏处理）、用户访问的时间、访问内容等多项属性的记录，并针对其中的各个属性进行说明见下表。

属性名称	属性说明	属性名称	属性说明
realIP	真实ip	fullURLId	网址类型
realAreacode	地区编号	hostname	源地址名
userAgent	浏览器代理	pageTitle	网页标题
userOS	用户浏览器类型	pageTitleCategoryId	标题类型ID
userId	用户ID	pageTitleCategoryName	标题类型名称
clientId	客户端ID	pageTitleKw	标题类型关键字
timestamp	时间戳	fullReferrer	入口源
timestamp_format	标准化时间	fullReferrerURL	入口网址
pagePath	路径	organicKeyword	搜索关键字
ymd	年月日	source	搜索源
fullURL	网址		

分析方法与过程

总体流程：



分析方法与过程

第1步：数据获取

- 网站每天的访问量有数千万次，非常巨大。为了便于初步分析，选取最近一段时间（2015-02-01~2015-04-29）广州市地区的所有用户访问的详细记录作为原始数据集，总共837450条记录。其中包括用户号、访问时间、来源网站、访问页面、页面标题、来源网页、标签、网页类别、关键词等。

第2步：数据探索与预处理

- 对数据进行多维度分析，用户访问时间，用户访问内容，流失用户分析以及用户分群等分析。
- 对数据进行预处理，包含数据去重，数据删选，数据分类等处理过程。
- 以用户访问html后缀的网页为关键条件，对数据进行处理。

1.数据清洗

2.属性规约

3.数据变换（属性构造、数据离散化）

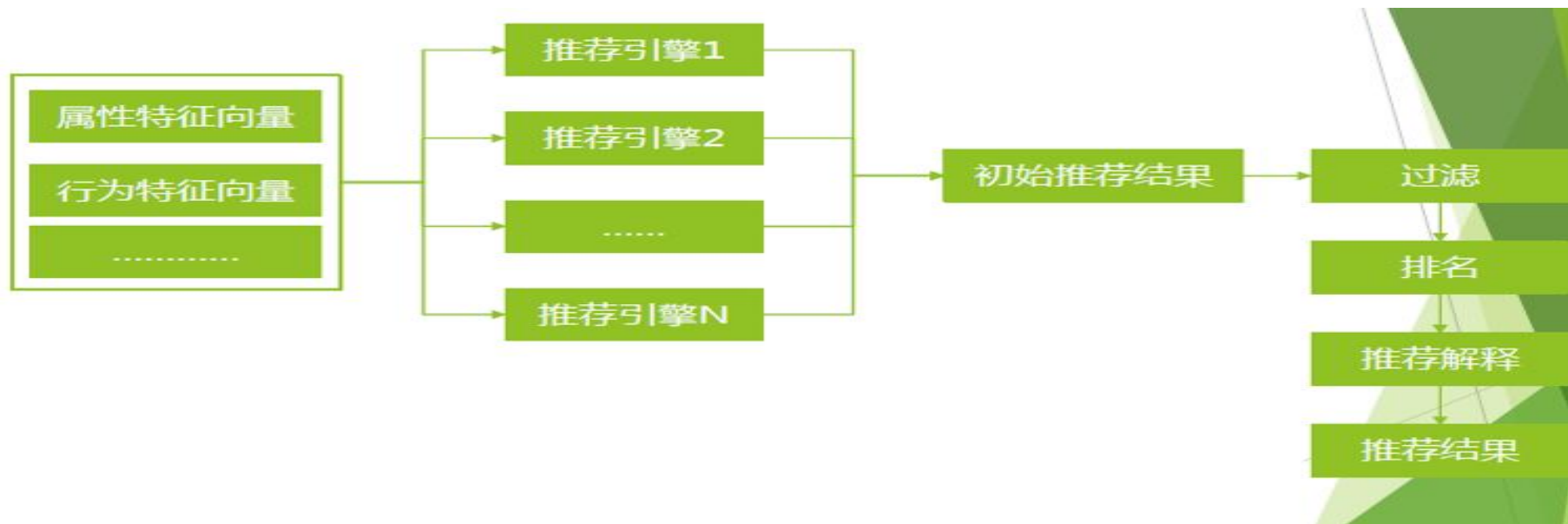


分析方法与过程

第3步：构建建模样本

对比多种推荐算法进行推荐，通过模型评价，得到比较好的智能推荐模型。通过模型对样本数据进行预测，获得推荐结果

第4步：构建模型



目录

1	电子商务网站智能推荐服务
2	智能推荐概述
3	协同过滤算法
4	上机实验
5	拓展思考



智能推荐概述

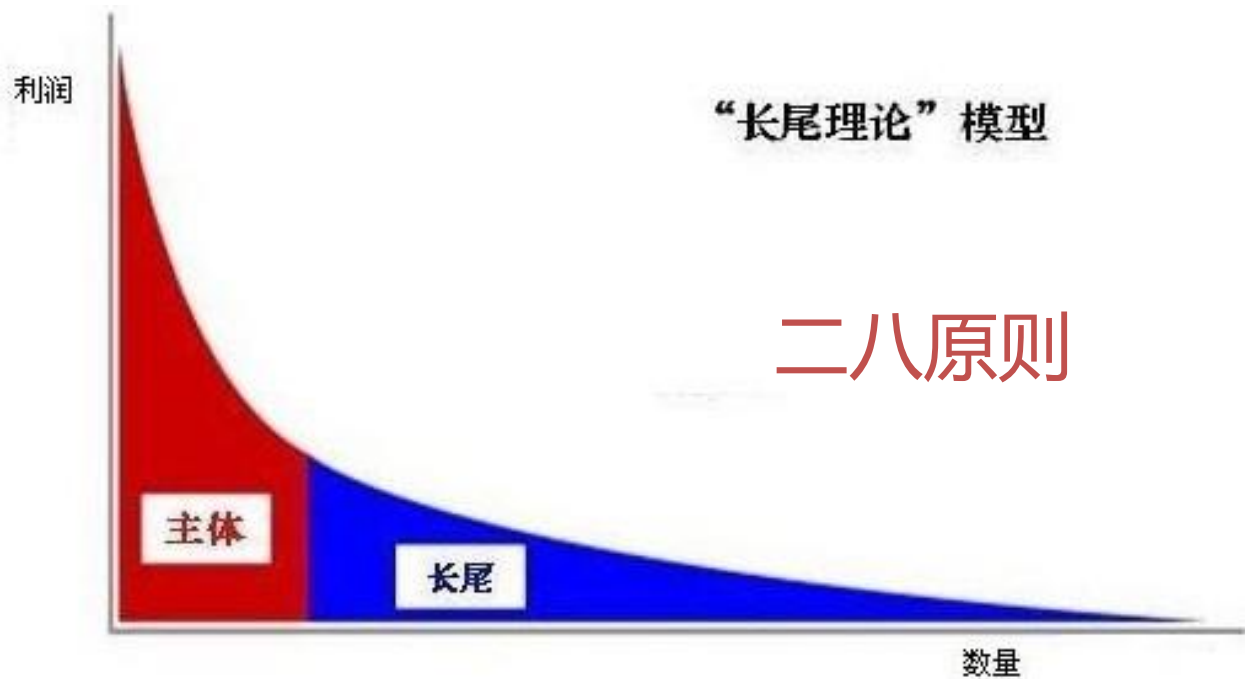
- 信息大爆炸时代来临，用户在面对大量的信息时无法从中迅速获得对自己真正有用的信息。
- 传统的搜索系统，需要用户提供明确需求，从用户提供的需求信息出发，继而给用户展现信息，无法针对不同用户的兴趣爱好提供相应地信息反馈服务。
- 推荐系统，相比于搜索系统，不需要用户提供明确需求，便可以为每一个用户实现个性化的推荐结果，让每个用户更便捷的获取信息。它是根据用户的兴趣特点和购买行为，向用户推荐用户感兴趣的信息和商品。



智能推荐概述

内容发现

- 信息过载 - 大数据时代
- 搜索引擎
 - 1. 你已经知道你要寻找什么
 - 2. 你愿意花时间去找
- 推荐引擎
 - 从长尾中意外地推荐你想要的物品
 - 用户不知道这个物品的存在
 - 预测用户多大程度喜欢某个物品



个性化/小众商品



智能推荐概述

推荐本质：通过一定的方式将用户和物品联系起来



智能推荐概述

智能推荐算法

基于关联规则的推荐算法

基于物品的协同过滤算法

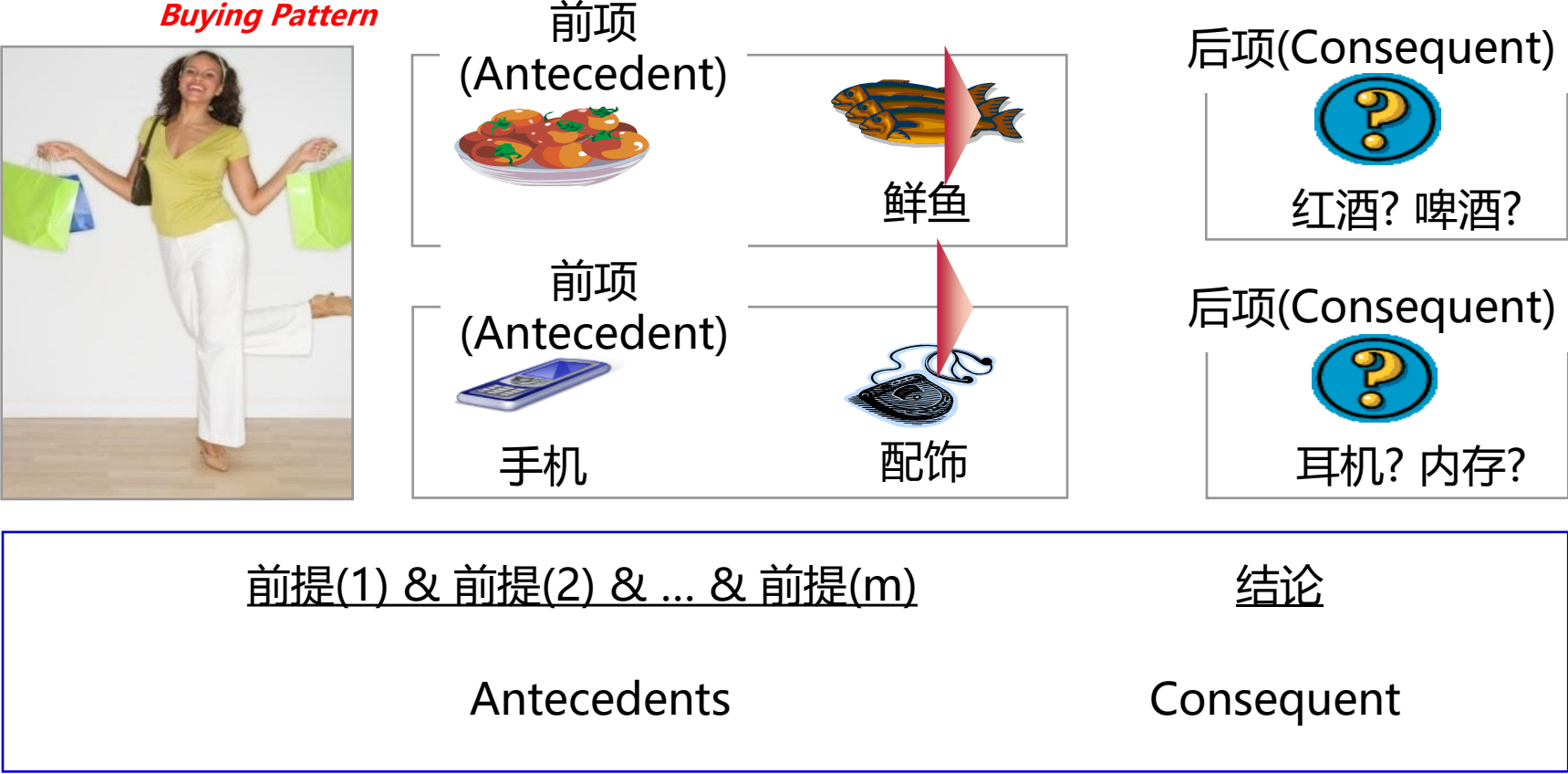
基于用户的协同过滤算法

基于内容的推荐算法



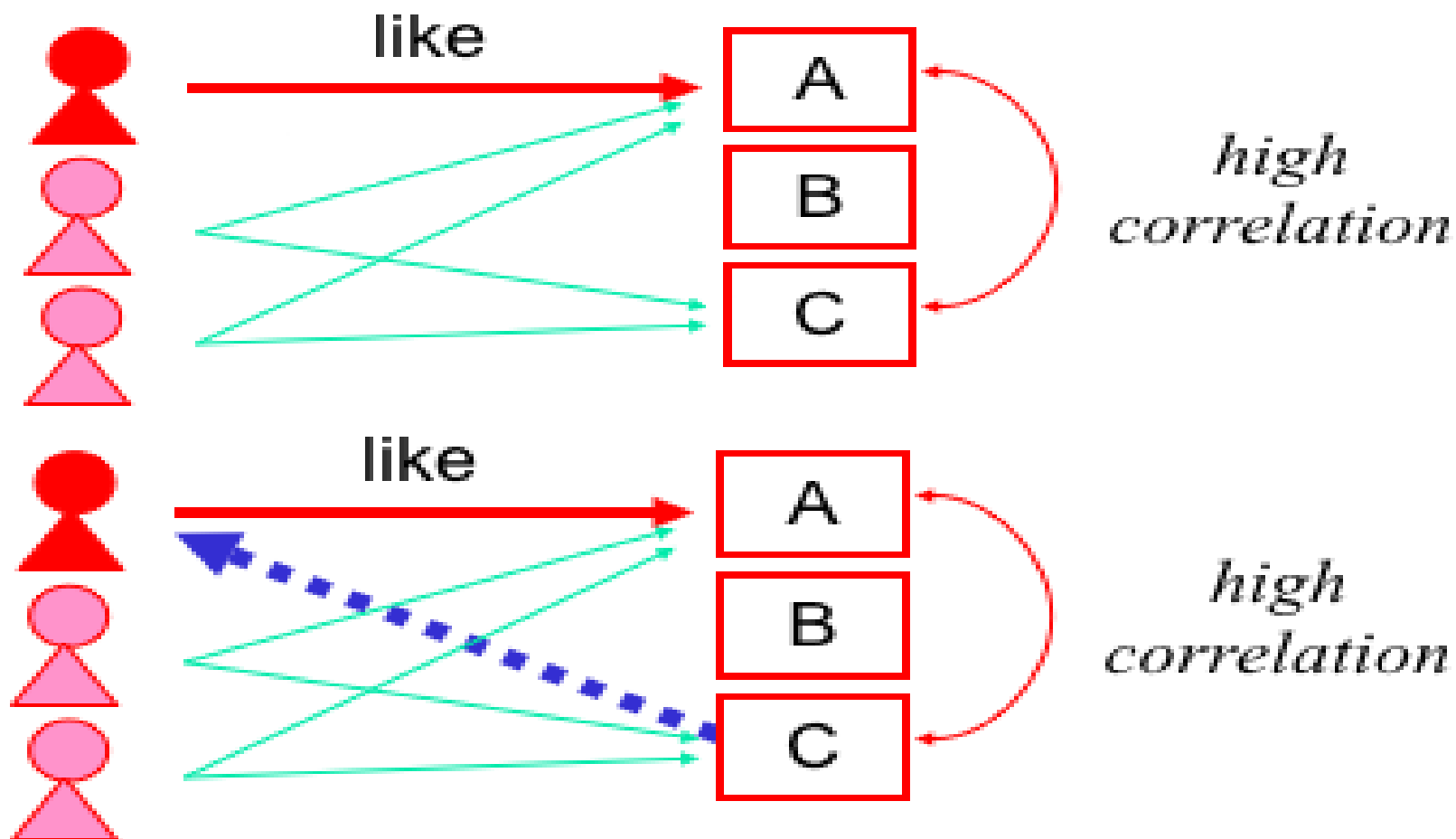
智能推荐概述

基于关联规则的推荐



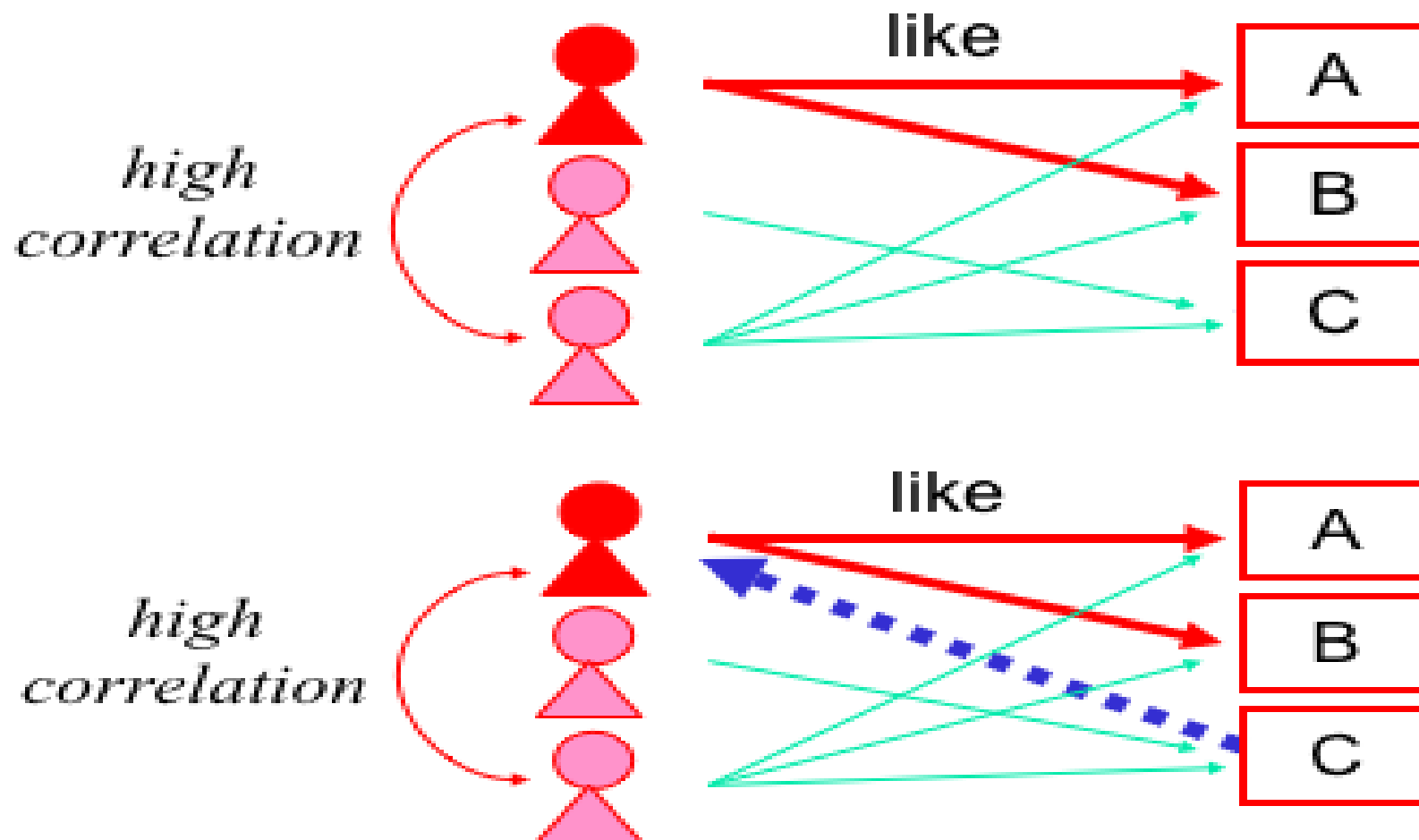
智能推荐概述

基于物品的协同过滤算法



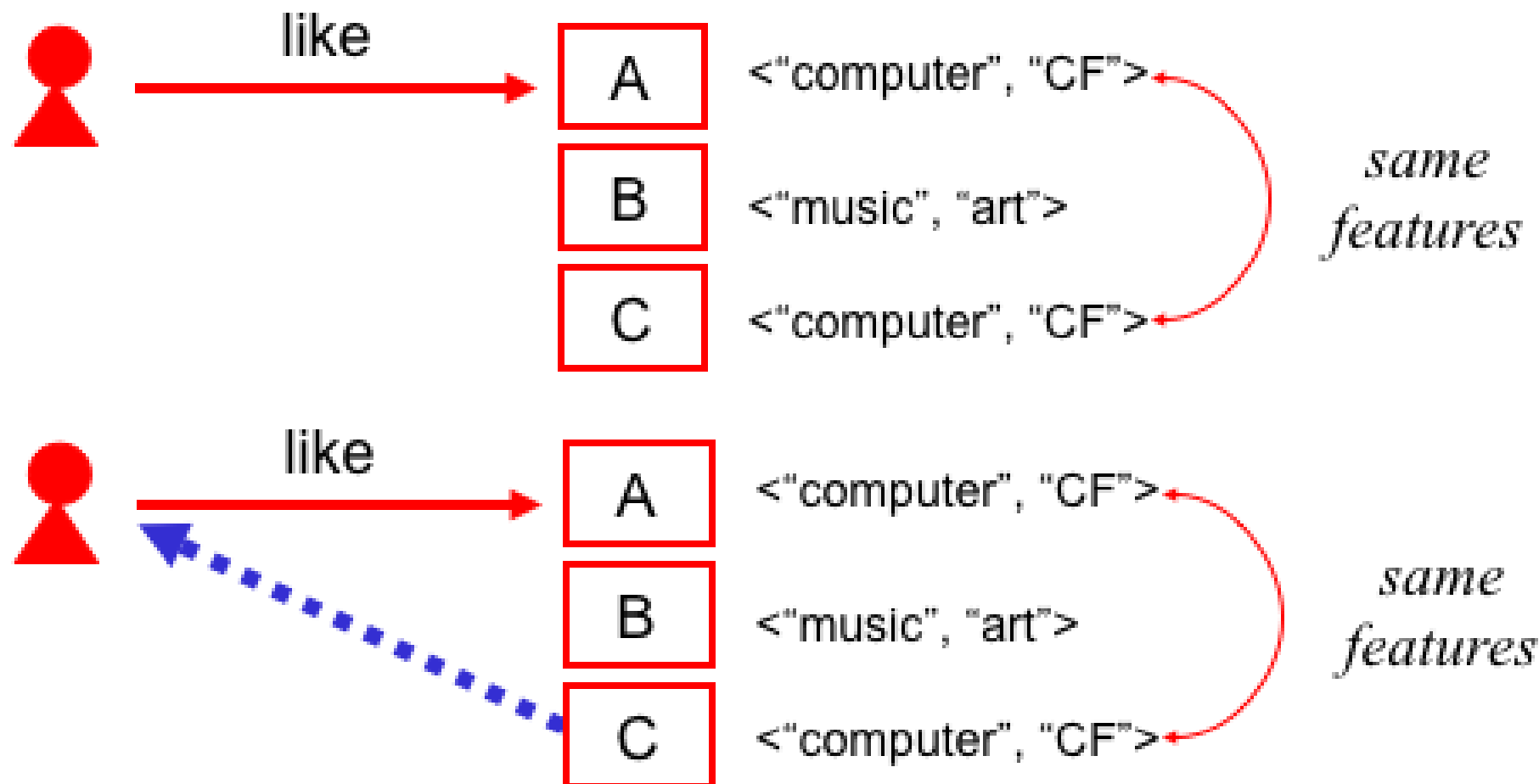
智能推荐概述

基于用户的协同过滤算法



智能推荐概述

基于内容的推荐算法



目录

1	电子商务网站智能推荐服务
2	智能推荐概述
3	协同过滤算法
4	上机实验
5	拓展思考



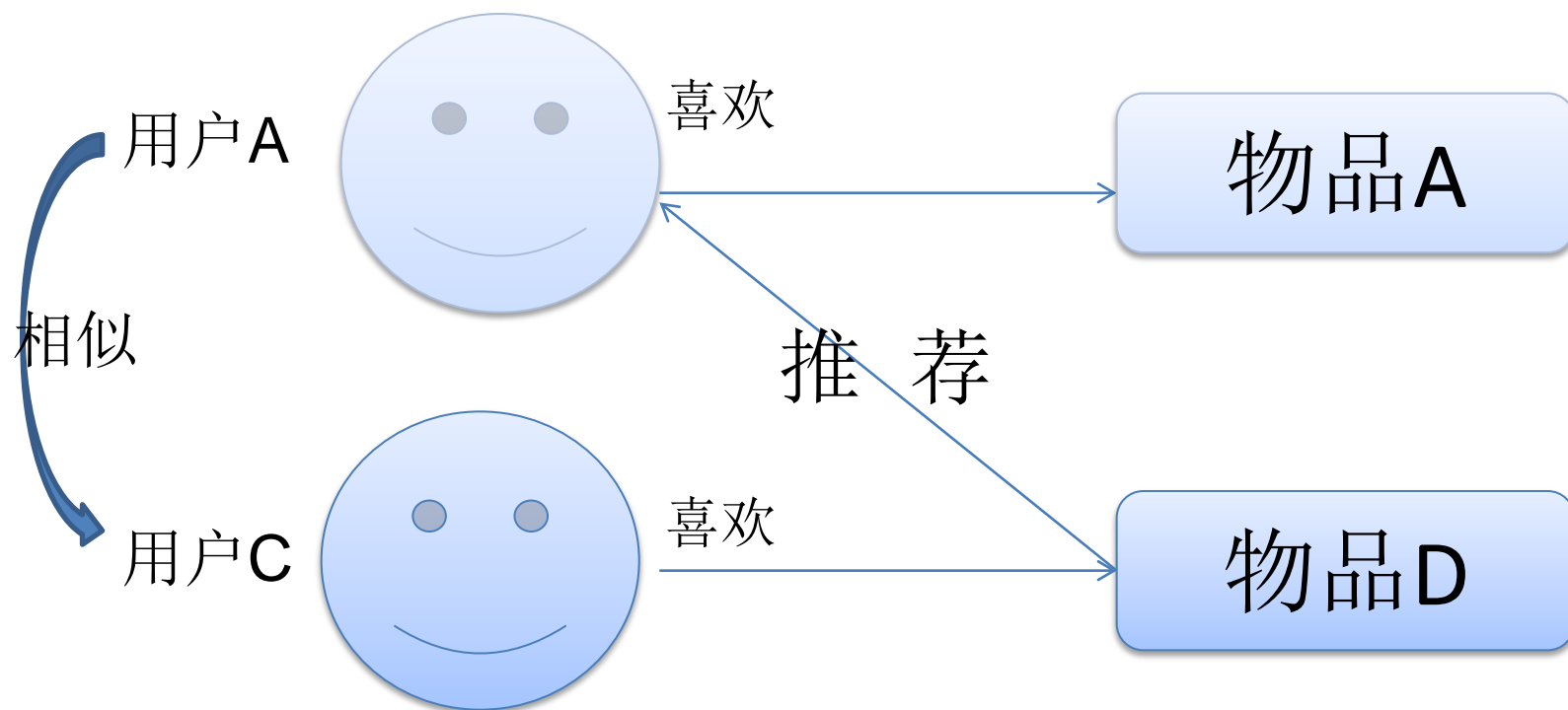
基于用户的协同过滤

- 智能推荐的方法有很多，常见的推荐技术主要分为：**基于用户的协同过滤推荐（UBCF）**和**基于物品的协同过滤推荐（IBCF）**。
- 基于用户的协同过滤的基本思想相当简单，基于用户对物品的偏好找到相邻邻居用户，然后将邻居用户喜欢的推荐给当前用户，为具有相同或相似的价值观、思想观、知识水平和兴趣偏好的用户，其对信息的需求也是相似的。
- 计算上，就是将一个用户对所有物品的偏好作为一个向量来计算用户之间的相似度，找到 K 邻居后，根据邻居的相似度权重以及他们对物品的偏好，预测当前用户没有偏好的未涉及物品，计算得到一个排序的物品列表作为推荐。



基于用户的协同过滤

- 下图 给出了一个例子，对于用户 A，根据用户的历史偏好，这里只计算得到一个邻居 - 用户 C，然后将用户 C 喜欢的物品 D 推荐给用户 A。



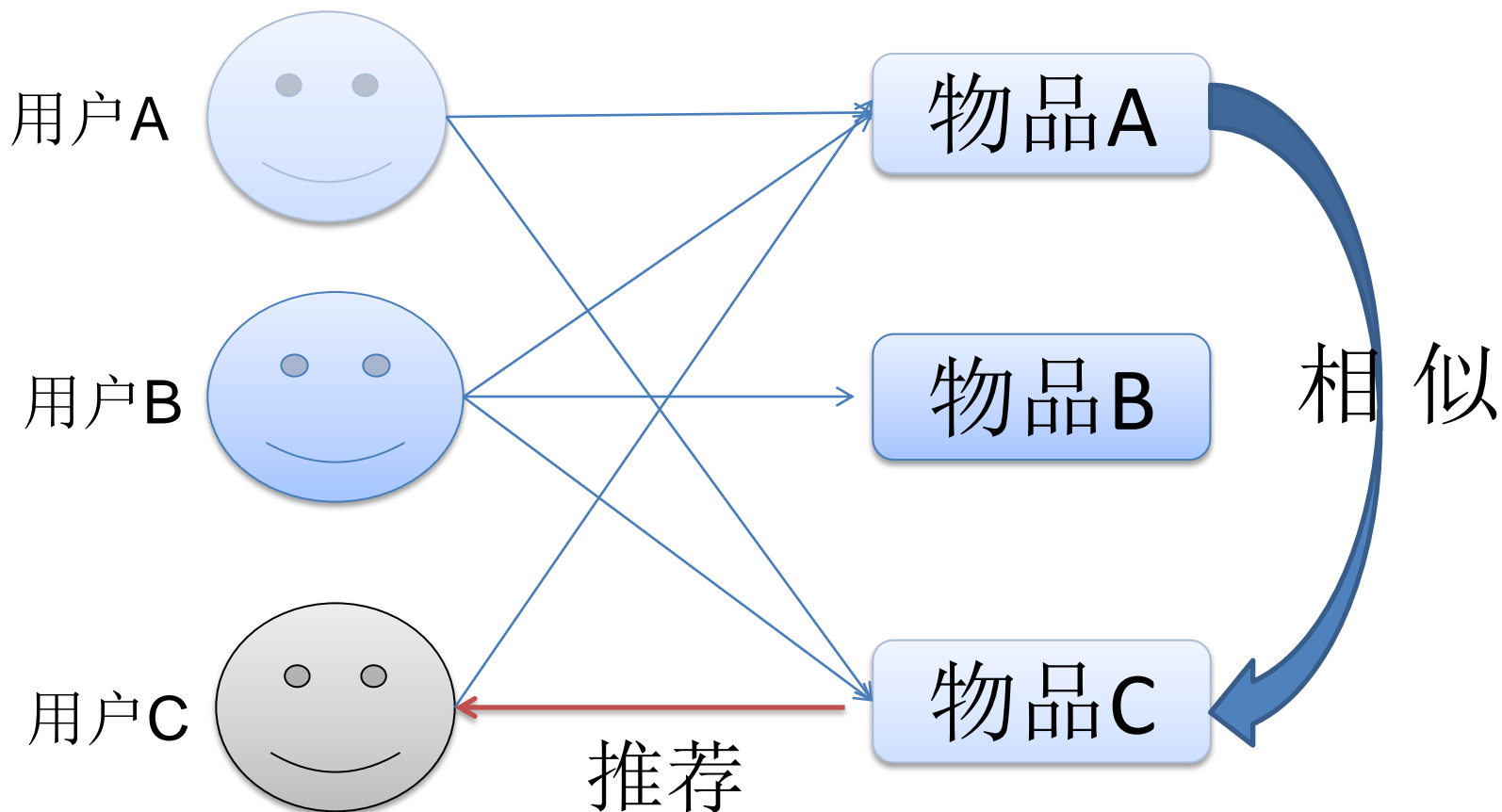
基于物品的协同过滤

- 基于物品的协同过滤的原理和基于用户的协同过滤类似，只是在计算邻居时采用物品本身，而不是从用户的角度，即基于用户对物品的偏好找到相似的物品，然后根据用户的历史偏好，推荐相似的物品给他。
- 从计算的角度看，就是将所有用户对某个物品的偏好作为一个向量来计算物品之间的相似度，得到物品的相似物品后，根据用户历史的偏好预测当前用户还没有表示偏好的物品，计算得到一个排序的物品列表作为推荐。



基于物品的协同过滤

- 下图给出了一个例子，对于物品 A，根据所有用户的历史偏好，喜欢物品 A 的用户都喜欢物品 C，得出物品 A 和物品 C 比较相似，而用户 C 喜欢物品 A，那么可以推断出用户 C 可能也喜欢物品 C。



基于用户的协同过滤

- 以电影评分数据为例，实现基于用户的协同过滤算法第一个重要的步骤就是计算用户之间的相似度。而计算相似度，建立相关系数矩阵目前主要分为以下几种方法。

a) 皮尔逊相关系数

- 皮尔逊相关系数一般用于计算两个定距变量间联系的紧密程度，它的取值在 $[-1, +1]$ 之间。用数学公式表示，皮尔森相关系数等于两个变量的协方差除于两个变量的标准差。计算公式如下所示：

$$s(X, Y) = \frac{Cov(X, Y)}{\sigma_X \sigma_Y}$$

- 由于皮尔逊相关系数描述的是两组数据变化移动的趋势，所以在基于用户的协同过滤系统中，经常使用。描述用户购买或评分变化的趋势，若趋势相近则皮尔逊系数趋近于1，也就是我们认为相似的用户。



基于用户的协同过滤

a) 皮尔逊相关系数

设 $(X_1, Y_1), \dots, (X_n, Y_n)$ 为从样本总体 $F(x, y)$ 中抽取的样本，则

$$E(X) = \bar{X} = \frac{1}{n} \sum_{i=1}^n X_i, \quad D(X) = S_X^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

$$E(Y) = \bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i, \quad D(Y) = S_Y^2 = \frac{1}{n-1} \sum_{i=1}^n (Y_i - \bar{Y})^2$$

$$\text{Cov}(X, Y) = S_{XY} = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y}) = E(XY) - E(X)E(Y)$$

相关系数和协方差的关系：
$$\rho_{XY} = \frac{S_{XY}}{\sqrt{S_X^2} \cdot \sqrt{S_Y^2}} = \frac{S_{XY}}{\sqrt{D(X)} \sqrt{D(Y)}} = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}$$



基于用户的协同过滤

a) 皮尔逊相关系数

$$\begin{aligned} Cov(X, Y) &= E[(X - E(X))(Y - E(Y))] \\ &= E(XY) - 2E(Y)E(X) - E(X)(Y) \\ &= E(XY) - E(X)E(Y) \end{aligned}$$

其中，X和Y不是相互独立，即它们之间存在着一定的关系。

协方差和方差的关系：

$$D(X + Y) = D(X) + D(Y) + 2Cov(X, Y)$$

$$D(X - Y) = D(X) + D(Y) - 2Cov(X, Y)$$



基于用户的协同过滤

b) 基于欧几里德距离的相似度余弦相似度

- 欧几里德距离计算相似度是所有相似度计算里面最简单、最易理解的方法
- 计算出来的欧几里德距离是一个大于0的数，为了使其更能体现用户之间的相似度，可以把它规约到(0, 1]之间，最终得到如下计算公式

$$s(X, Y) = \frac{1}{1 + \sum \sqrt{(X_i - Y_i)^2}}$$

- 只要至少有一个共同评分项，就能用欧几里德距离计算相似度；如果没有共同评分项，那么欧几里德距离也就失去了作用。
- 其实照常理理解，如果没有共同评分项，那么意味着这两个用户或物品根本不相似。



基于用户的协同过滤

c) 余弦相似度

- 余弦相似度用向量空间中两个向量夹角的余弦值作为衡量两个个体间差异的大小。余弦相似度更加注重两个向量在方向上的差异，而非距离或长度上。计算公式如下所示：

$$s(X, Y) = \cos \theta = \frac{\vec{x} * \vec{y}}{\|x\| * \|y\|}$$

例如：

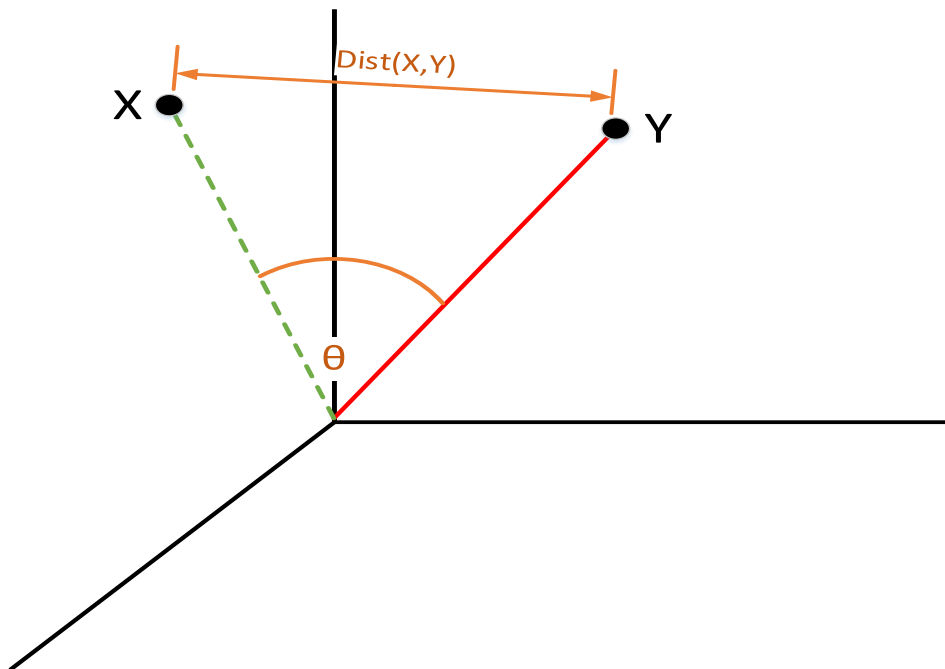
我\是\数据\分析师。

我\是\数据\挖掘专家。



基于用户的协同过滤

- 从图上可以看出距离度量衡量的是空间各点间的绝对距离，跟各个点所在的位置坐标（即个体特征维度的数值）直接相关。
- 如果保持X点的位置不变，Y点朝原方向远离坐标轴原点，那么这个时候余弦相似度是保持不变的，因为夹角不变，而X、Y两点的距离显然在发生改变，这就是欧氏距离和余弦相似度的不同之处。



基于用户的协同过滤

d) 杰卡德系数

- Jaccard相似系数 (Jaccard similarity coefficient) 用于比较有限样本集之间的相似性与差异性。Jaccard系数值越大，样本相似度越高。

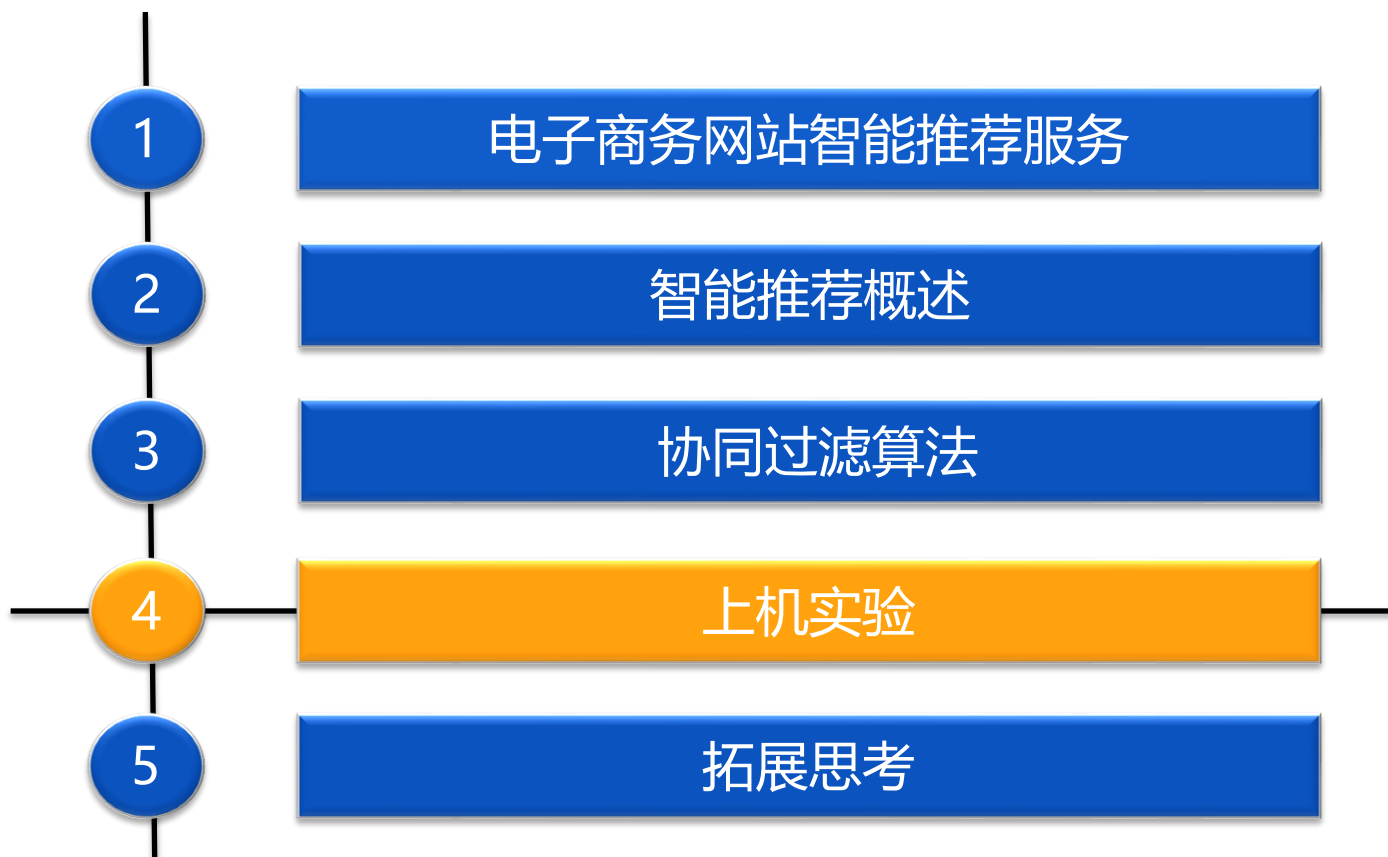
$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}$$

主要应用场景：

- 比较文本相似度，用于文本查重与去重；
- 计算对象间距离，用于数据聚类等



目录



上机实验

- 下面通过电影推荐的例子演示基于用户的协同过滤算法在Python中的实现。
- 现在影视已经成为大众喜爱的休闲娱乐的方式之一，合理的个性化电影推荐一方面能够促进电影行业的发展，另一方面也可以让大众数量众多的电影中迅速得到自己想要的电影，从而做到两全齐美。甚至更近一步，可以明确市场走向，对后续电影的类型导向等起到重要作用。
- 现有的部分电影评分数据如下表：

Index	Claudia Puig	Gene Seymour	Jack Matthews	Lisa Rose	Michael Phillips	Mick LaSalle	Toby
Just My Luck	3	1.5	0	3	0	2	0
Lady in the Water	0	3	0	2.5	2.5	3	0
Snakes on a Plane	3.5	3.5	4	3.5	3	0	4.5
Superman Returns	0	0	5	3.5	3.5	3	4
The Night Listener	4.5	3	3	0	4	3	0
You, Me and Dupree	2.5	0	3.5	2.5	0	2	1

上机实验

- # 计算不同电影的相似度，将数据规范化为[0 , 1]
- `corr = mdata.corr(method = 'pearson')`
- `corr = 0.5+corr*0.5`
- # 或者
- `mcors = np.corrcoef(mdata, rowvar=0)`
- `mcors = 0.5+mcors*0.5`
- `mcors = pd.DataFrame(mcors, columns=mdata.columns, index=mdata.columns)`



上机实验

- # 计算每个用户的每个项目的分数
- def cal_score(matrix,mcors,item,user):
- totscore = 0
- totsims = 0
- score = 0
- if pd.isnull(matrix[item][user]) or matrix[item][user]==0:
- for mitem in matrix.columns:
- if matrix[mitem][user]==0:
- continue
- else:
- totscore += matrix[mitem][user]*mcors[item][mitem]
- totsims += mcors[item][mitem]
- score = totscore/totsims
- else:
- score = matrix[item][user]
- return score



上机实验

- # 计算socre矩阵
- # score_matrix: 得分矩阵:不同用户的电影得分矩阵
- def cal_matscore(matrix,mcors):
- score_matrix = pd.DataFrame(np.zeros(matrix.shape) , columns=matrix.columns, index=matrix.index)
- for mitem in score_matrix.columns:
- for muser in score_matrix.index:
- score_matrix[mitem][muser] = cal_score(matrix,mcors,mitem,muser)
- return score_matrix



上机实验

- # 根据得分矩阵给出推荐
- def recommend(matrix,score_matrix,user,n):
- user_ratings = matrix.ix[user]
- not_rated_item = user_ratings[user_ratings==0]
- recom_items = {}
- for item in not_rated_item.index:
- recom_items[item] = score_matrix[item][user]
- recom_items = pd.Series(recom_items)
- recom_items = recom_items.sort_values(ascending=False)
- return recom_items[:n]

- # 开始推荐
- score_matrix = cal_matscore(mdata,mcors)
- for i in range(10):
- user = input(str(i)+'please input the name of user:')
- print(recommend(mdata,score_matrix,user,2))



目录

1	电子商务网站智能推荐服务
2	智能推荐概述
3	协同过滤算法
4	上机实验
5	拓展思考



拓展思考

- 协同过滤技术是目前推荐系统中最成功和应用最广泛的技术，在理论研究和实践中都取得了快速的发展，相对于其它的推荐技术，由于协同过滤通过“推荐对象”的特征信息来了解用户的兴趣，并能够发现用户的潜在兴趣，具备较高的个性化程度，因此协同过滤技术受到越来越多研究者的关注，并广泛应用于在电子商务推荐领域。
- 协同过滤也称为社会过滤，它计算用户间偏好的相似性，在相似用户的基础上自动的为目标用户进行过滤和筛选，其基本思想为具有相同或相似的价值观、思想观、知识水平和兴趣偏好的用户，其对信息的需求也是相似的。
- 因此相对于传统的推荐方法，协同过滤技术体现出的一个显著的优势是能够推荐一些难以进行内容分析的项目，比如信息质量、个人品味等抽象的资源对象。另外协同过滤技术能够有效的使用其他兴趣相似用户的评价信息，从而利用较少的用户反馈，加快了个性化学习的速度，同时利于发现用户的隐藏兴趣。



拓展思考

- 下面通过个性化的电影推荐的例子演示基于用户的协同过滤算法在Python中的实现。
- 现在影视已经成为大众喜爱的休闲娱乐的方式之一，合理的个性化电影推荐一方面能够促进电影行业的发展，另一方面也可以让大众数量众多的电影中迅速得到自己想要的电影，从而做到两全齐美。甚至更近一步，可以明确市场走向，对后续电影的类型导向等起到重要作用。
- 现有的部分电影评分数据如下表：

用户ID	电影ID	电影评分	时间标签
1	1	5	874965758
1	2	3	876893171
1	3	4	878542960
1	4	3	876893119
1	5	3	889751712
1	6	4	875071561
1	7	1	875072484
...

拓展思考

➤ 在Python中实现基于用户的协同过滤推荐系统首先计算用户之间的相关系数。实现代码如下所示：

- #使用基于UBCF算法对电影进行推荐
- import pandas as pd
- if __name__ == "__main__":
- print("\n-----使用基于UBCF算法对电影进行推荐 运行中... -----\n")
- traindata = pd.read_csv('../data/u1.base',sep='\t', header=None,index_col=None)
- testdata = pd.read_csv('../data/u1.test',sep='\t', header=None,index_col=None)
- traindata.drop(3,axis=1, inplace=True)
- testdata.drop(3,axis=1, inplace=True)
- traindata.rename(columns={0:'userid',1:'movid',2:'rat'}, inplace=True)
- testdata.rename(columns={0:'userid',1:'movid',2:'rat'}, inplace=True)
- traindf=traindata.pivot(index='userid', columns='movid', values='rat')
- testdf=testdata.pivot(index='userid', columns='movid', values='rat')
- traindf.rename(index={i:'usr%d'%(i) for i in traindf.index}, inplace=True)
- traindf.rename(columns={i:'mov%d'%(i) for i in traindf.columns}, inplace=True)
- testdf.rename(index={i:'usr%d'%(i) for i in testdf.index}, inplace=True)
- testdf.rename(columns={i:'mov%d'%(i) for i in testdf.columns}, inplace=True)
- userdf=traindf.loc[testdf.index]
- #获取预测评分和推荐列表
- trainrats,trainrecomm=recomm(traindf,userdf)



拓展思考

➤ 输出结果

usr1([u'mov1290', u'mov1354', u'mov1678'], dtype='object', name=u'movid'),
usr2([u'mov1491', u'mov1354', u'mov1371'], dtype='object', name=u'movid'),
usr3([u'mov1304', u'mov1621', u'mov1678'], dtype='object', name=u'movid'),
usr4([u'mov1502', u'mov1659', u'mov1304'], dtype='object', name=u'movid'),
usr5([u'mov1304', u'mov1621', u'mov1472'], dtype='object', name=u'movid'),
usr6([u'mov1618', u'mov1671', u'mov1357'], dtype='object', name=u'movid'),
usr7([u'mov1472', u'mov1467', u'mov1374'], dtype='object', name=u'movid'),
usr8([u'mov1659', u'mov1316', u'mov1494'], dtype='object', name=u'movid'),
usr9([u'mov1621', u'mov1304', u'mov1491'], dtype='object', name=u'movid'),
usr10([u'mov1486', u'mov1494', u'mov437'], dtype='object', name=u'movid'),
usr11([u'mov1659', u'mov1654', u'mov1626'], dtype='object', name=u'movid'),
usr12([u'mov1659', u'mov1618', u'mov1661'], dtype='object', name=u'movid'),
usr13([u'mov1486', u'mov1494', u'mov1662'], dtype='object', name=u'movid'),
usr14([u'mov1661', u'mov1308', u'mov1671'], dtype='object', name=u'movid'),
usr15([u'mov1626', u'mov1671', u'mov1678'], dtype='object', name=u'movid'),
usr16([u'mov1618', u'mov1486', u'mov1494'], dtype='object', name=u'movid'),
usr17([u'mov1316', u'mov1621', u'mov1304'], dtype='object', name=u'movid'),
usr18([u'mov1618', u'mov1654', u'mov1626'], dtype='object', name=u'movid'),
usr19([u'mov1316', u'mov1661', u'mov1275'], dtype='object', name=u'movid'),
usr20([u'mov1659', u'mov1292', u'mov1304'], dtype='object', name=u'movid'),
Total: 80000rows



拓展思考

- 对输出结果进行解释：其中最前端格式为“usr+整数”字符串代表用户编号，“[]”内的字符串代表三部电影的编号，dtype为类型，name为字段名。
- 整体代表意思是，根据算法得出对用户usr1推荐他并未看过的三部电影，编号为：mov1290，mov1354，u'mov1678。





大数据成就未来



Thank you!

泰迪科技 : www.tipdm.com
热线电话 : 40068-40020

