



# 新零售推荐系统：从算法到应用

尼奥

2018.05.26

# C 目录 ONTENTS

1

推荐系统概述

2

新零售中的推荐

3

推荐算法的应用

# 信息的价值

$$\text{信息的价值} = (I * V) S$$

I: 信息量

V: 信息传递速度

S: 信息共享范围



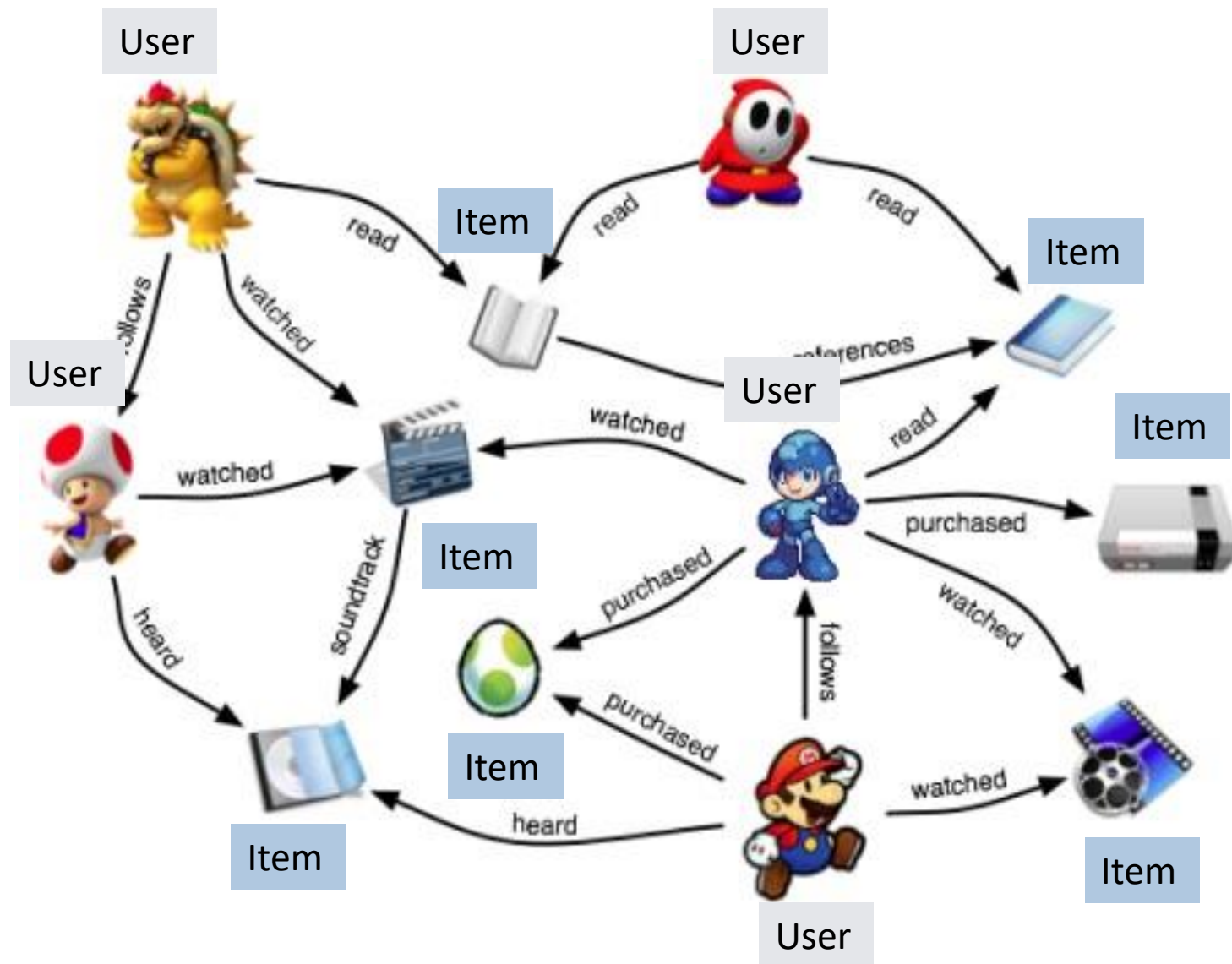
# 推荐系统概述

- 什么是推荐系统？

- Predict preference or rating of an 'item' from a user ——wiki

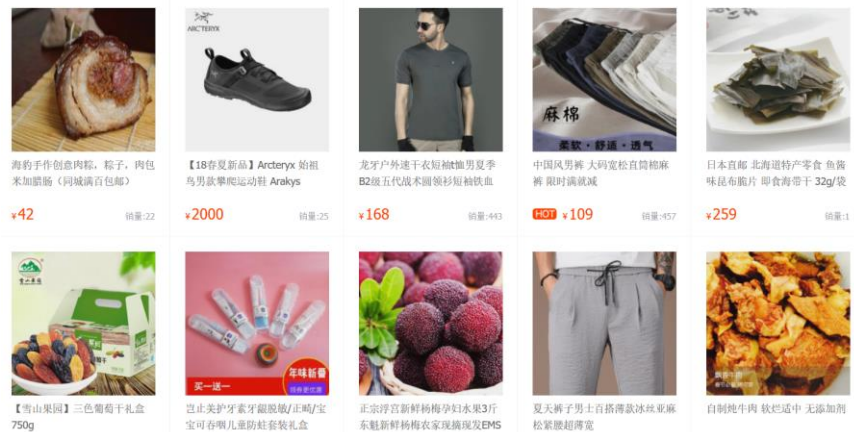
- 千人千面

- 增加点击
- 发掘长尾



# 常见推荐系统应用场景

猜你喜欢



海豹手作创意肉粽, 粽子, 肉包米加腊肠 (同城满包邮) +42 销量:22

【18春夏新品】Arcteryx 始祖鸟新款攀爬运动鞋 Aralys +2000 销量:25

龙牙户外速干衣短袖男夏季B2级五代战术面料短袖快血 +168 销量:443

中国风男裤 大码宽松直筒棉麻裤 限时满减 +109 销量:457

日本直邮 北海道特产零食 鱼贝味昆布脆片 即食海带干 32g/袋 +259 销量:1

【香山果园】三色葡萄干礼盒 750g

巴止美护牙素牙齦凝胶(止痛)宝宝可吞服儿童防蛀套装礼盒

正宗淳宜新鲜杨梅孕妇水果3斤 东魁新鲜杨梅农家现摘现发EMS

夏天裤子男白色百搭冰丝亚麻松紧腰超薄宽

自制纯牛肉 软烂适中 无添加剂

电商平台

即时热点



进击, 中国军队 03:07 3.5万次播放

这是我的临阵演说 04:31 2.1万次播放

张靓颖婚内多次与男闺蜜暧昧? 两人暧昧期间被扒 11:43 26.1万次播放

佟丽娅拍戏太保守 白敬亭在枕头上藏菜刀 01:13 1.5万次播放

三星堆文明起源之谜 究竟与外星人是否有关系? 13:01 1,152次播放

龙袍御用御膳! 古代奢侈品云锦比黄金还贵? 10:48 1,132次播放

优酷懂你



马苏红毯走6分钟被吐槽 刘雨欣称接受张檬道歉 09:46 90.20万次播放

探秘档案: 黄河清淤挖出两龙棺材? 黄河透明棺材之谜 06:20 24.33万次播放

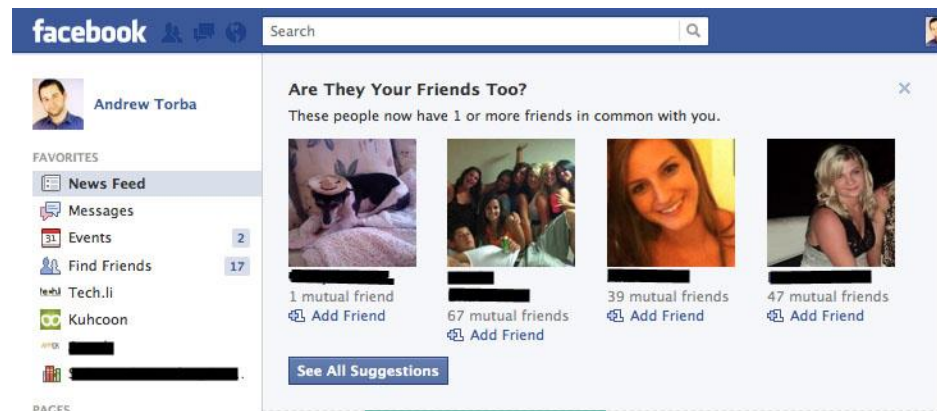
愤怒! 3名外中国游戏路边女孩 00:52 8.98万次播放

唐唐说电影: 最催眠的猛片和猪队友玩密室逃脱 12:11 104.2万次播放

范冰冰被老外认成李冰冰 谢娜未修图遭网友diss 09:14 34.94万次播放

《白夜追凶 第二季》宣传预告片 00:11 141.4万次播放

影音平台



facebook Search

Andrew Torba

Are They Your Friends Too? These people now have 1 or more friends in common with you.

1 mutual friend Add Friend

67 mutual friends Add Friend

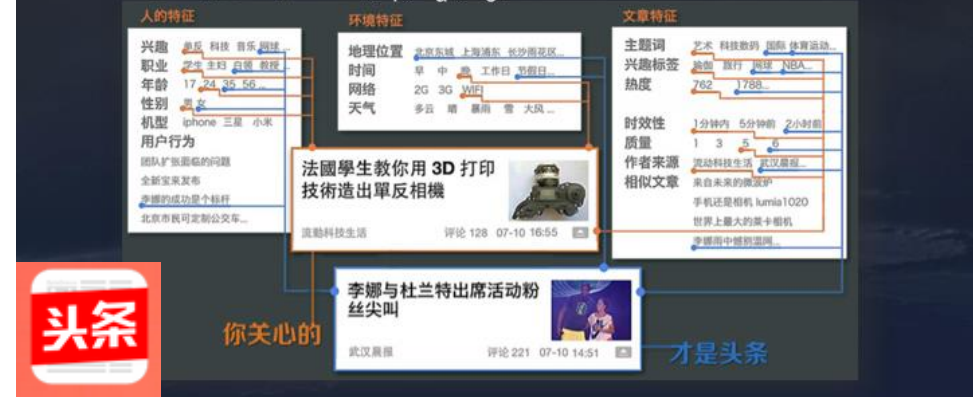
39 mutual friends Add Friend

47 mutual friends Add Friend

See All Suggestions

社交平台

资讯推荐系统本质上要解决用户、环境和资讯的匹配:  $y = F(x_i, x_u, x_c)$



人的特征: 兴趣 (音乐, 科技, 游戏), 职业 (学生, 宝妈, 医生, 教师), 年龄 (17, 24, 36, 56), 性别 (男, 女), 机型 (iphone, 三星, 小米), 用户行为 (团队扩张面临的问题, 全新宝来发布, 李娜的成功是个怪样, 北京市民可定制公交车)

环境特征: 地理位置 (北京东城, 上海浦东, 长沙雨花区), 时间 (早, 中, 晚, 工作日, 节假日), 网络 (2G, 3G, WiFi), 天气 (多云, 晴, 暴雨, 雪, 大风...)

文章特征: 主题词 (艺术, 科技, 游戏, 财经, 体育, 运动), 兴趣标签 (旅行, 篮球, NBA), 热度 (762, 1788), 时效性 (1分钟内, 5分钟内, 2小时内), 质量 (1, 3, 5), 作者来源 (流酷科技生活, 武汉晨报, 来自未来的微浪乎), 相似文章 (手机还是相机 lumia1020, 世界上最大的最牛相机, 李娜雨中摔跤落泪)

法国学生教你用3D打印技术造出单反相机

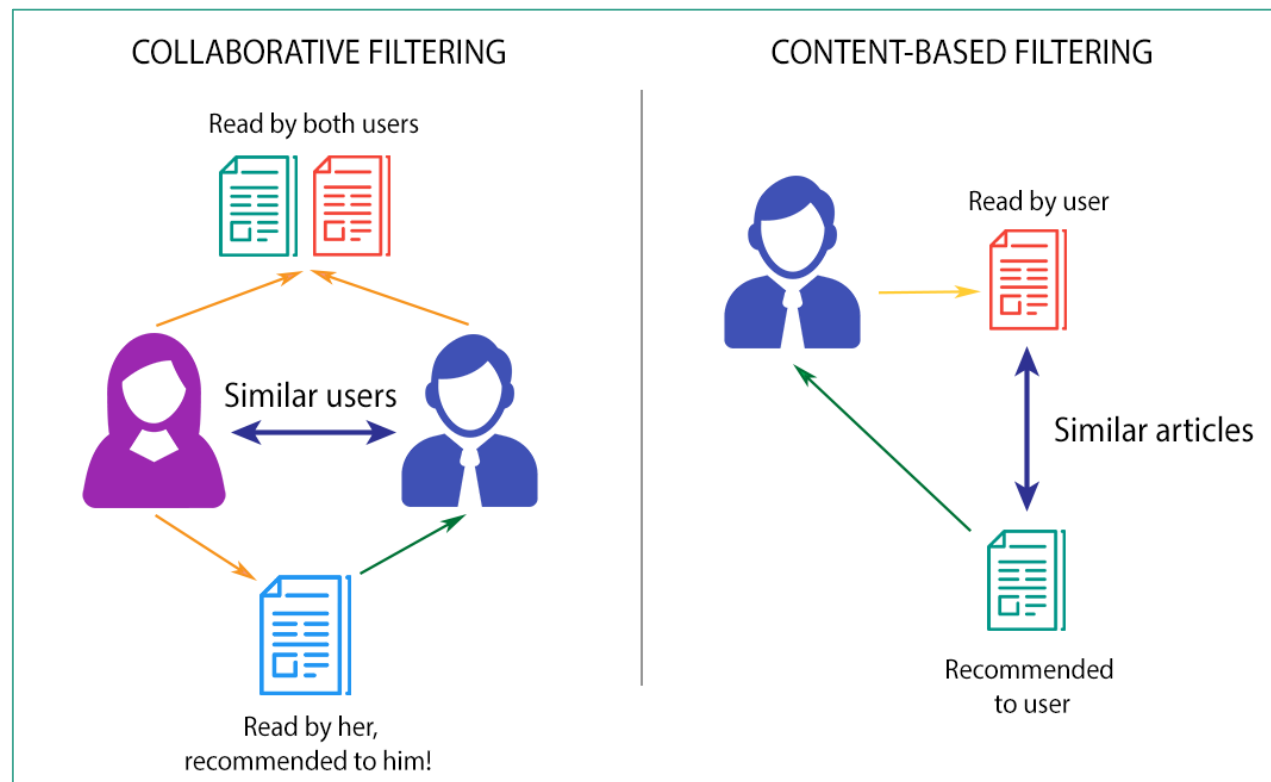
李娜与杜兰特出席活动粉丝尖叫

你关心的 才是头条

资讯平台

# 常用推荐算法

- 基于内容：历史记录
- 协同过滤
  - 基于用户（资讯）
  - 基于物品（电商）
- 关联规则：非监督学习
- 基于效用：关注点
- 基于知识：交互





# 协同过滤：UserCF与ItemCF

UserCF:

计算用户相似度  
需要维护用户信息矩阵  
新用户的冷启动  
更社会化

ItemCF:

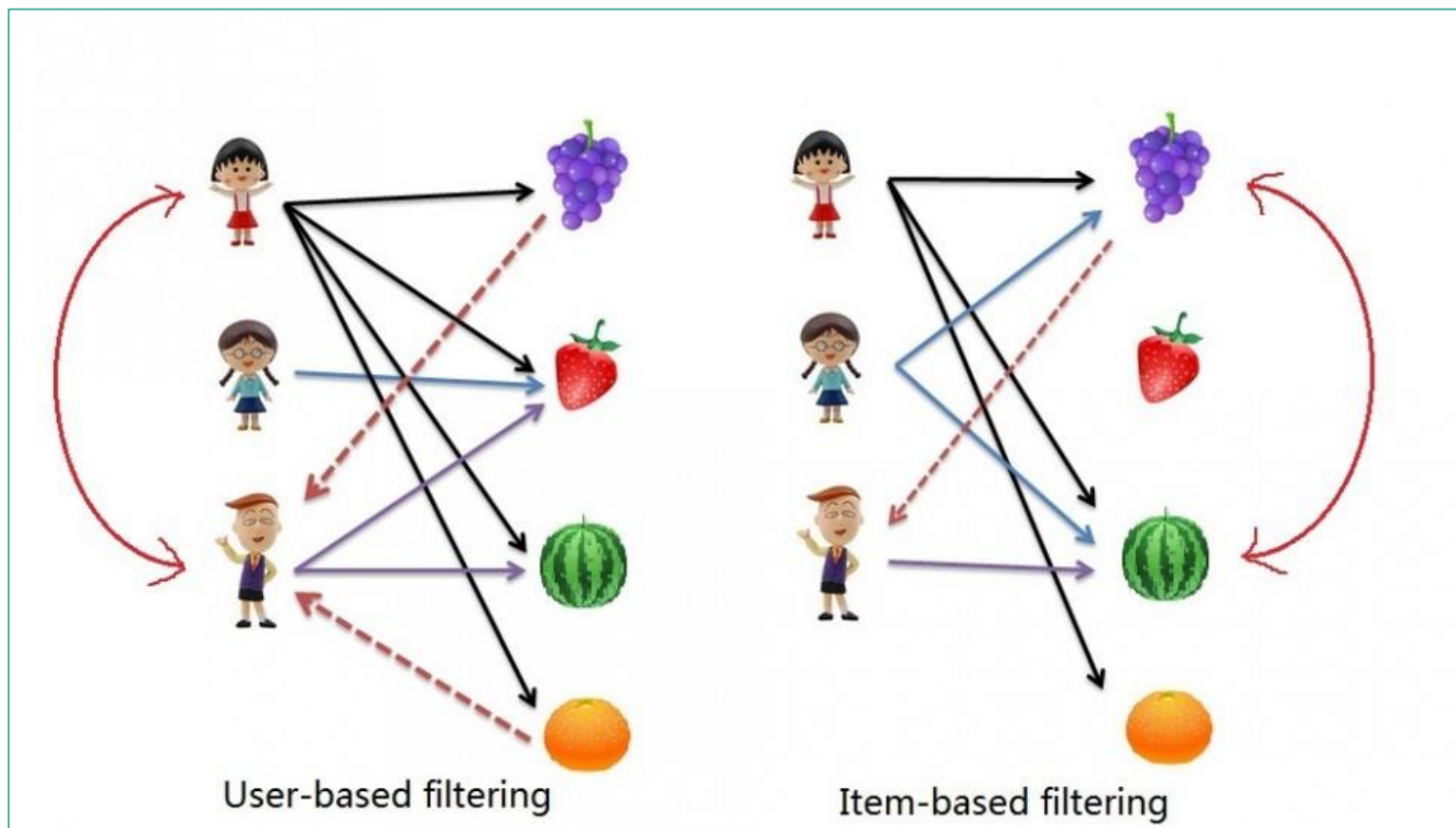
计算物品相似度  
需要维护物品的信息矩阵  
新物品的冷启动  
更个性化

UserCF-IIF

$$w_{uv} = \frac{\sum_{i \in N(u) \cap N(v)} \frac{1}{\log 1 + |N(i)|}}{\sqrt{|N(u)| |N(v)|}}$$

ItemCF-IUF

$$w_{ij} = \frac{\sum_{u \in N(i) \cap N(j)} \frac{1}{\log 1 + |N(u)|}}{\sqrt{|N(i)| |N(j)|}}$$



# 协同过滤：评分预测算法

$$r_{ui} \approx \mathbf{q}_i^* \mathbf{p}_u.$$

$$\min_{\mathbf{q}^*, \mathbf{p}^*} \sum_{(u,i) \in \mathcal{K}} (r_{ui} - \mathbf{q}_i^* \mathbf{p}_u)^2 + \lambda (\|\mathbf{q}_i\|^2 + \|\mathbf{p}_u\|^2).$$

$$e_{ui} = r_{ui} - \mathbf{q}_i^* \mathbf{p}_u$$

$$\begin{aligned} \mathbf{q}_i &\leftarrow \mathbf{q}_i + \gamma (e_{ui} \mathbf{p}_u - \lambda \mathbf{q}_i) \\ \mathbf{p}_u &\leftarrow \mathbf{p}_u + \gamma (e_{ui} \mathbf{q}_i - \lambda \mathbf{p}_u). \end{aligned}$$

ALS

$$b_{ui} = \mu + b_u + b_i$$

$$\min_{b_*} \sum_{(u,i) \in \mathcal{K}} (r_{ui} - \mu - b_u - b_i)^2 + \lambda_1 \left( \sum_u b_u^2 + \sum_i b_i^2 \right)$$

$$r_{ui} = \mu + b_i + b_u + \mathbf{q}_i^* \mathbf{p}_u$$

$$\min_{p^*, q^*, b_*} \sum_{(u,i) \in \mathcal{K}} (r_{ui} - \mu - b_u - b_i - p_u^T q_i)^2 + \lambda_3 (\|p_u\|^2 + \|q_i\|^2 + b_u^2 + b_i^2)$$

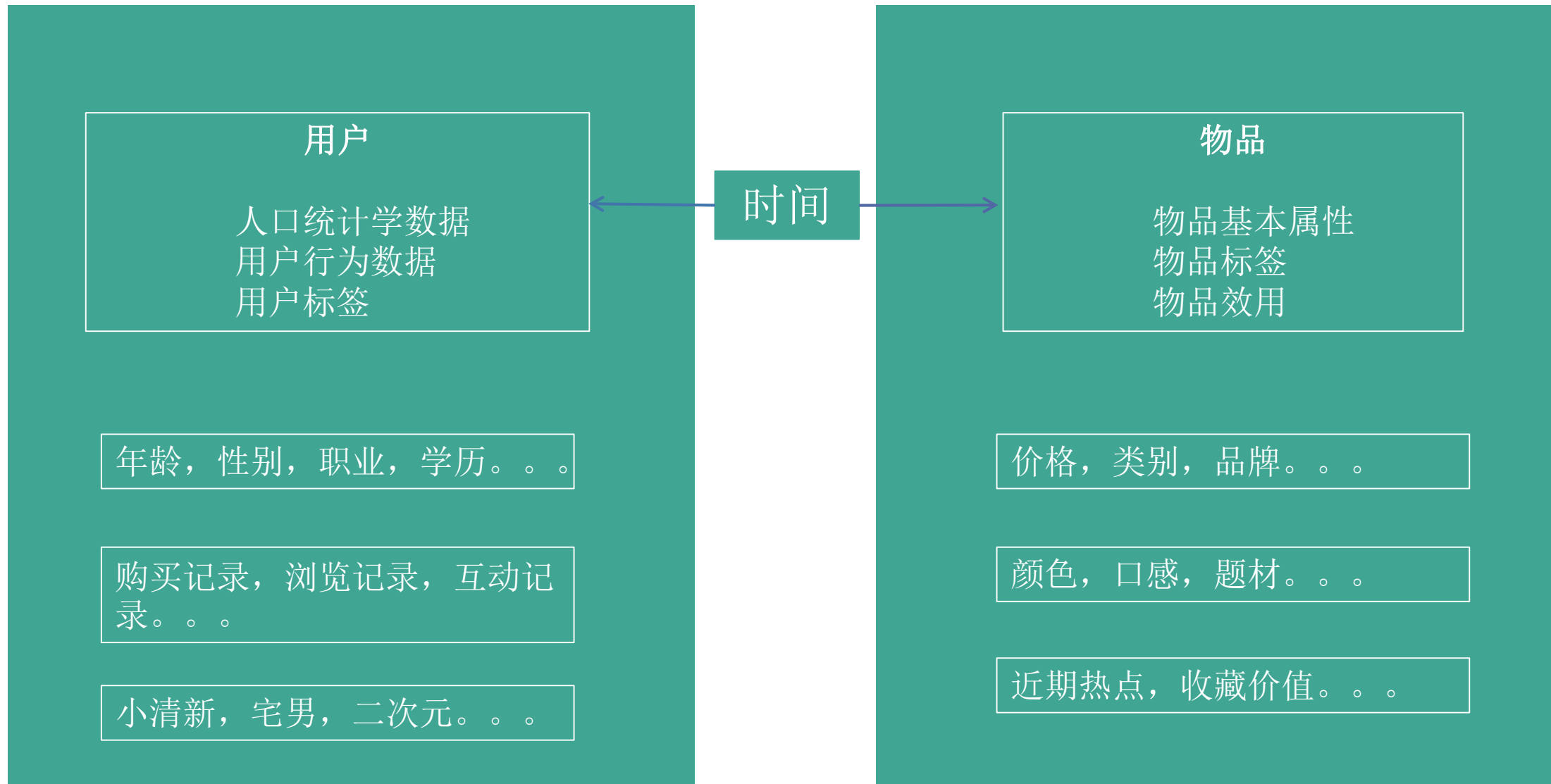
$$\hat{r}_{ui} = b_{ui} + q_i^T \left( p_u + |N(u)|^{-\frac{1}{2}} \sum_{j \in N(u)} y_j \right)$$

$$\min_{p, q, b} \sum_{u,i} (r_{ui} - \mu - b_u - b_i - q_i^T (p_u + |N(u)|^{-1/2} \sum_{j \in N(u)} y_j))^2 + \lambda (\|p_u\|^2 + \|q_i\|^2 + b_u^2 + b_i^2 + \sum_{j \in N(u)} \|y_j\|^2)$$

SVD++



# 推荐需要的数据



# 推荐结果的检验

准确率

$$\text{Precision} = \frac{\sum_{u \in U} R(u) \cap T(u)}{\sum_{u \in U} R(u)}$$

召回率

$$\text{Recall} = \frac{\sum_{u \in U} R(u) \cap T(u)}{\sum_{u \in U} T(u)}$$

$$F_{\beta} = \frac{(1 + \beta)^2 \text{Recall} \cdot \text{Precision}}{\beta^2 \cdot \text{Recall} + \text{Precision}}$$

覆盖率

$$\text{Coverage} = \frac{|\bigcup_{u \in U} R(u)|}{|I|}$$

多样性

$$\text{Diversity} = \frac{1}{|U|} \sum_{u \in U} \text{Diversity}(R(u))$$

新颖度/流行度

# 算法：机器学习的五大流派

符号主义：逻辑学/哲学； 逆向演绎

联结主义：神经科学； 反向传播

进化主义：进化生物学； 遗传算法

贝叶斯派：统计学； 概率推理

行为类推主义：心理学； 机器内核

推荐算法

## The Five Tribes of Machine Learning

Tribe	Origins	Master Algorithm
Symbolists	Logic, philosophy	Inverse deduction
Connectionists	Neuroscience	Backpropagation
Evolutionaries	Evolutionary biology	Genetic programming
Bayesians	Statistics	Probabilistic inference
Analogizers	Psychology	Kernel machines

<https://www.leiphone.com/news/201608/nBZ8goAlOaKEEYrQ.html?viewType=weixin>

# C 目录 ONTENTS

1

推荐系统概述

2

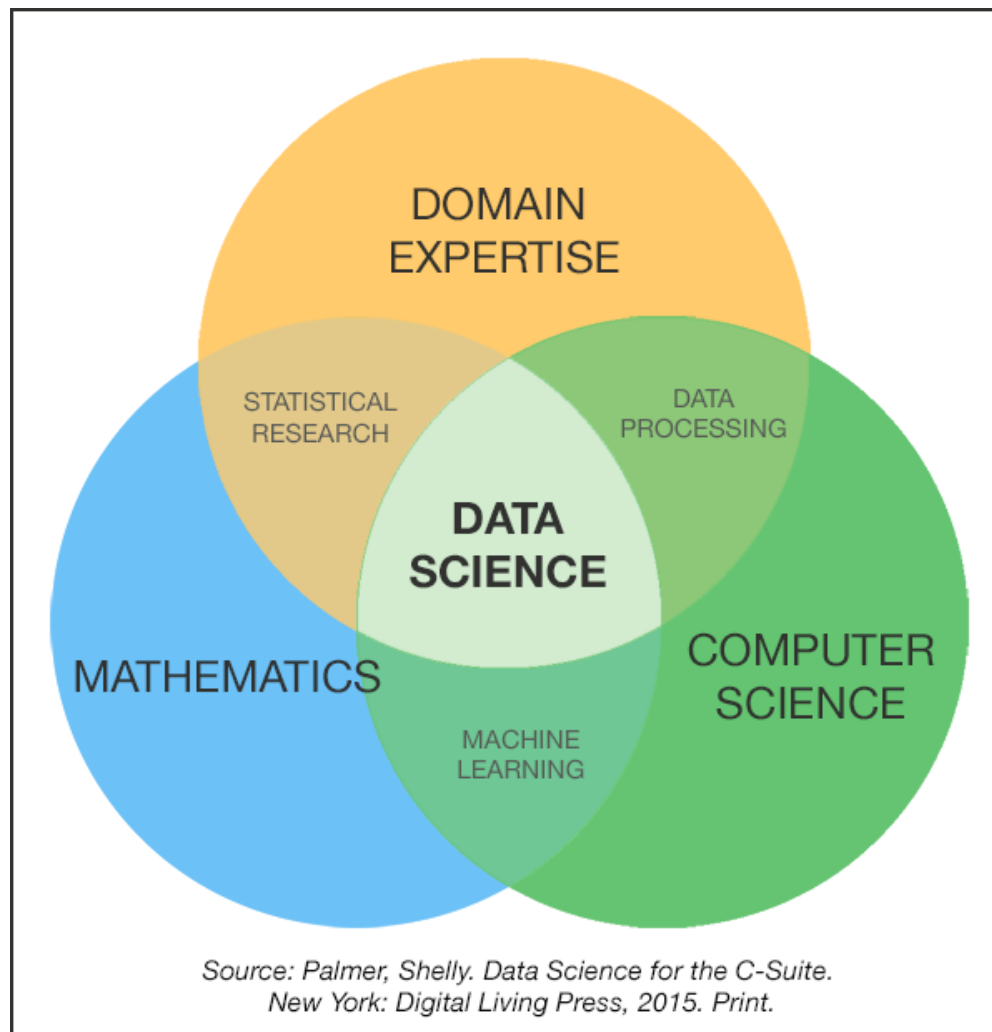
新零售中的推荐

3

推荐算法的应用

# 推荐系统：除了技术，还有业务

- 技术主导
- 业务主导



# 什么是新零售?

新零售 = 线上 + 线下 + 物流

- 盒马鲜生，超级物种
- 小米体验店，天猫小店
- 微信群营销

“

未来的十年、二十年，没有电子商务这一说，只有新零售这一说，也就是说线上线下和物流必须结合在一起，才能诞生真正的新零售

”





# 新零售中的推荐：与互联网推荐的区别

## 互联网中的推荐

Predict preference or rating of an 'item' from a user ——wiki

### 千人千面

- 增加点击
- 发掘长尾

## 新零售中的推荐

预测用户的喜好  
——线上与线下

### 千店千面

- 增加销量
- 信息共享



# 新零售中的推荐：与互联网推荐的区别

可用信息少：

行为信息  
物品信息

需求不同：

UserCF为主

业务性强：

筛选规则

反馈周期长：

结果检验不易

图书推荐 (更新时间: 2017-01-01, 每周更新一次)

书名	作者	出版日期	定价	ISBN	分类	近30天销量	门店库存	推荐采购量	推荐原因
<input type="checkbox"/> 深海危机/中国原创奇幻动物小说	唐池子	2015-12-01	16.0	9787534289637	中国儿童文学	3	0	10	相似门店热销
<input type="checkbox"/> 爱学拼音(上)/魔力铅笔	爱学系列...	2009-06-01	3.0	9787534252747	学前教育	15	0	30	相似门店热销
<input type="checkbox"/> 科学野战营(史无前例的发明)	纸上魔方	2016-08-01	20.0	9787534293665	少儿百科词典	45	5	10	相似门店热销
<input type="checkbox"/> 小学生生字卡(3上)	林彤	2015-06-01	10.5	9787534285868	小学语文	18	0	20	相似门店热销
<input type="checkbox"/> 三国演义/世界少年文学经典文库	罗贯中	2009-05-01	14.5	9787534253881	儿童文学	37	0	20	相似门店热销

全选

添加图书

# C 目录 ONTENTS

1

推荐系统概述

2

推荐算法简介

3

推荐算法的落地

推荐算法上线了，一切才刚刚开始。。。

强特征是永恒的话题

流行度 vs 覆盖率

结果计算速度太慢怎么办

相似度的第n+1种算法

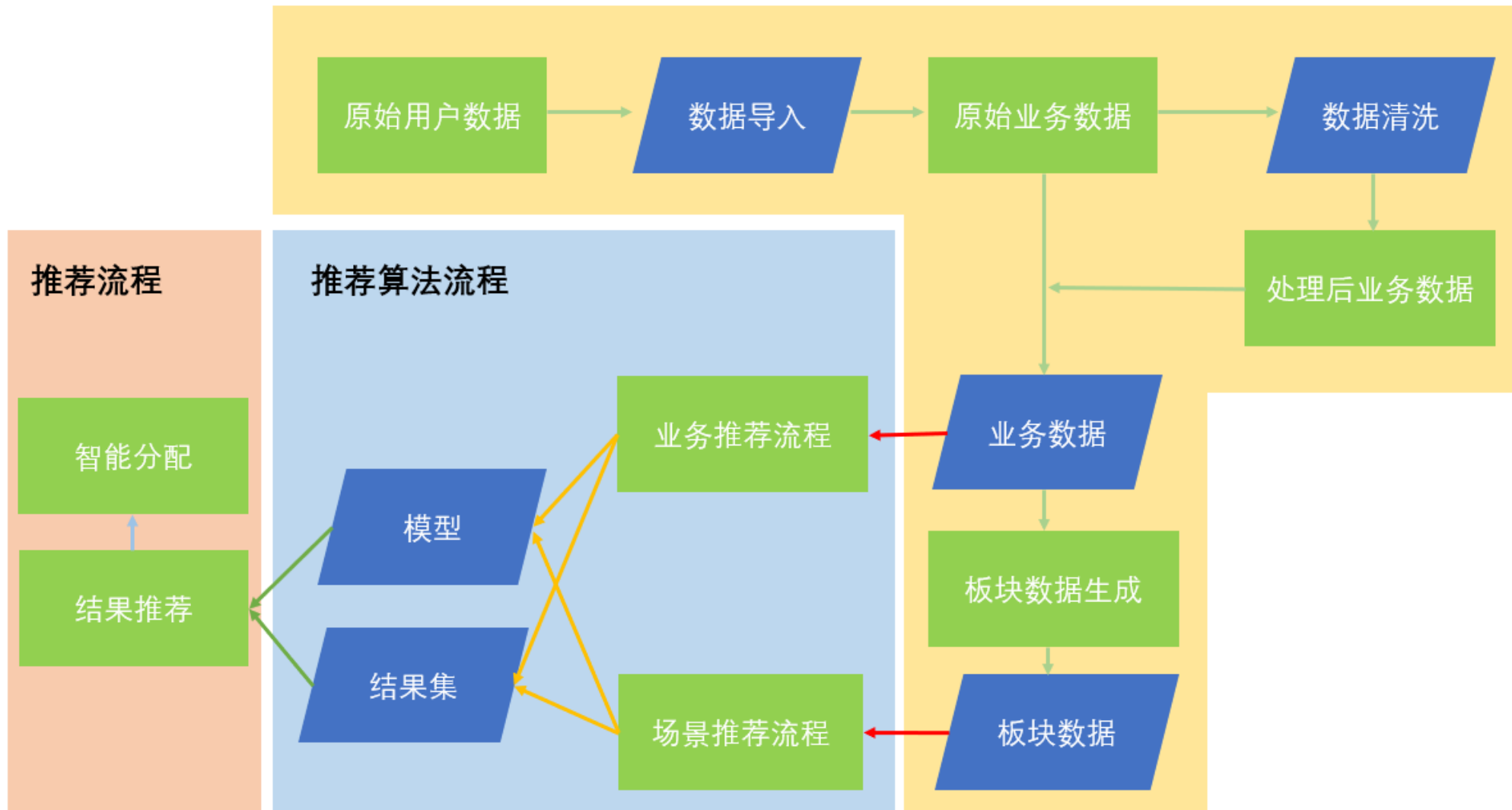
永远都有新想法的客户

机器验证 vs 人工验证

特征不仅会增加，还会消失。。。

# 算法架构

## 数据处理流程



## Netflix Awards \$1 Million Prize and Starts a New Contest

BY STEVE LOHR SEPTEMBER 21, 2009 10:15 AM



Jason Kempin/Getty Images Netflix prize winners, from left: Yehuda Koren, Martin Chabbert, Martin Piotte, Michael Jahrer, Andreas Toscher, Chris Volinsky and Robert Bell.

**Update | 1:45 p.m.** Adding details announced Monday about the extremely close finish to the contest.

And in fact, it was. Despite all the plaudits and case studies, Netflix announced this week that despite paying \$1 million dollars to a winning team of multinational researchers in 2009, they never bothered to implement their solution. Why? Because, according to Netflix the "additional accuracy gains that we measured did not seem to justify the engineering effort needed to bring them into a production environment."

## Robust De-anonymization of Large Datasets (How to Break Anonymity of the Netflix Prize Dataset)

Arvind Narayanan and Vitaly Shmatikov

The University of Texas at Austin

February 5, 2008

### Abstract

We present a new class of statistical de-anonymization attacks against high-dimensional micro-data, such as individual preferences, recommendations, transaction records and so on. Our techniques are robust to perturbation in the data and tolerate some mistakes in the adversary's background knowledge.

We apply our de-anonymization methodology to the Netflix Prize dataset, which contains anonymous movie ratings of 500,000 subscribers of Netflix, the world's largest online movie rental service. We demonstrate that an adversary who knows only a little bit about an individual subscriber can easily identify this subscriber's record in the dataset. Using the Internet Movie Database as the source of background knowledge, we successfully identified the Netflix records of known users, uncovering their apparent political preferences and other potentially sensitive information.



# 推荐在新零售中的应用：一些感想

- 1, 业务效果好的算法并不一定是复杂的算法
- 2, 做项目要比光看书的收获大得多
- 3, 推荐是个系统工程, 算法很重要, 但不是全部



# 数据智能 让未来变成现在

[niao@dtstack.com](mailto:niao@dtstack.com)



邮箱: [support@dtstack.com](mailto:support@dtstack.com)

地址: 浙江省杭州市紫霞街176号杭州互联网创新创业园2号楼8F

网站: [www.dtstack.com](http://www.dtstack.com)