推荐系统初探

A Survey of Recommender Systems

张永锋 清华大学计算机系

2012.4.9

Outline

- 推荐系统概述
  - 什么是推荐(个性化推荐)
  - 个性化推荐的发展历史
  - 个性化推荐问题的输入输出
  - 个性化问题的形式化
  - 推荐系统的两大核心问题
- 个性化推荐方法分类及典型算法
  - 多种分类方法
  - 典型算法介绍与对比
- 推荐算法的评价方法和指标
  - 评价方法
  - 评价指标
- 典型商业推荐系统浅析
  - 典型商业推荐系统
- 面临的问题及发展方向
  - 推荐系统目前面临的问题
  - 推荐系统潜在的发展方向
- Demo
  - 数据集介绍
  - demo

推荐系统概述

- 推荐系统概述
  - 什么是(个性化)推荐
  - 个性化推荐的发展历史
  - 个性化推荐问题的输入输出
  - 个性化推荐问题的形式化
  - 推荐系统的两大核心问题

什么是推荐系统

- 推荐系统:
  - Recommender System
  - 预测用户对某个他未曾“使用”过的物品(item)的喜好程度
  - 物品item：电影、书籍、音乐、新闻……
  - 一般所说的“推荐系统”是指“个性化推荐系统”，即不同人根据具体情况不同可以获得不同的、有针对性的推荐

推荐问题的发展历史

- 推荐问题本身追溯久远

- 1994, Minnesota, GroupLens研究组论文[1]
  - 提出“协同过滤”的概念
  - 推荐问题的形式化
  - 影响深远(An Open Architecture)

- GroupLens : user-based collaborative filtering netnews Recommendation System
  - Item-based
  - Matrix Factorization
  - Other non-CF algorithms
  - Hybrid Methods

- http://www.grouplens.org/
- recommender systems
- online communities

- [1]GroupLens: An Open Architecture for Collaborative Filtering of Netnews, CSCW 1994. 3034 Refs

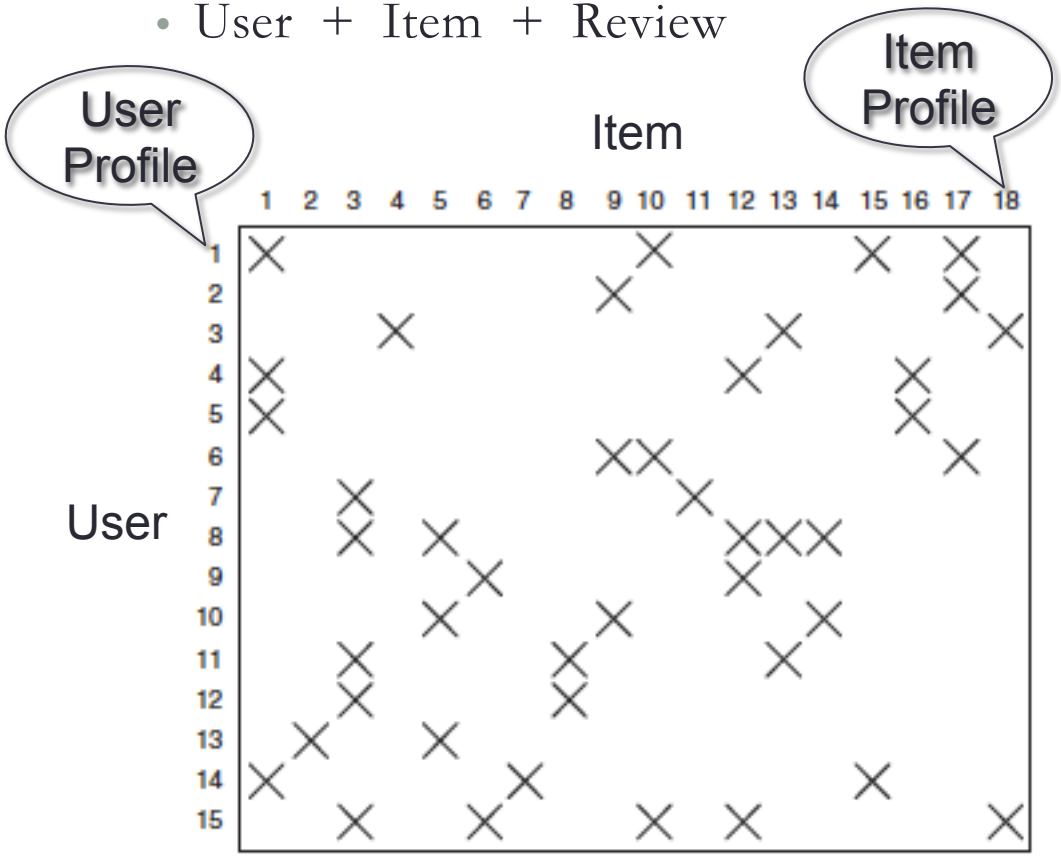
推荐问题的发展历史(cont.)

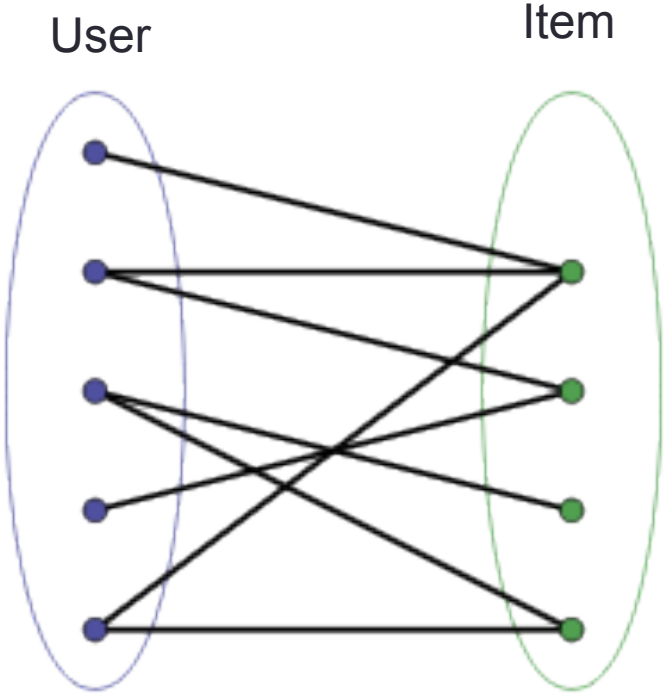
- 目前已广泛集成到很多商业应用系统中
  - 尤其是网络购物平台中

- Amazon：
  - Amazon网络书城的推荐算法每年贡献30个百分点的创收
- Forrester:
  - 电子商务网站留意到推荐信息的顾客，约1/3会依据推荐购买商品

推荐系统的输入

- 推荐系统的输入
  - User + Item + Review





推荐系统的输入(cont.)

- Item & Item Profile
  - 电影：类别、导演、主演、国家、……
  - 新闻：标题、本文、关键词、时间、……
- User & User Profile
  - 描述一个user的“个性”
  - 两种构建User Profile的方式
    - 与Item Profile类似，如性别、年龄、国别、年收入、活跃时间……
      - 难以与Item建立具体的联系
      - 隐私问题
      - 很少直接使用
    - 利用Item Profile构建User Profile[2]
      - Personalized IR related

- [2] Adaptive Web Search based on User Profile Constructed without Any Effort from Users, WWW2004.

推荐系统的输入(cont.)

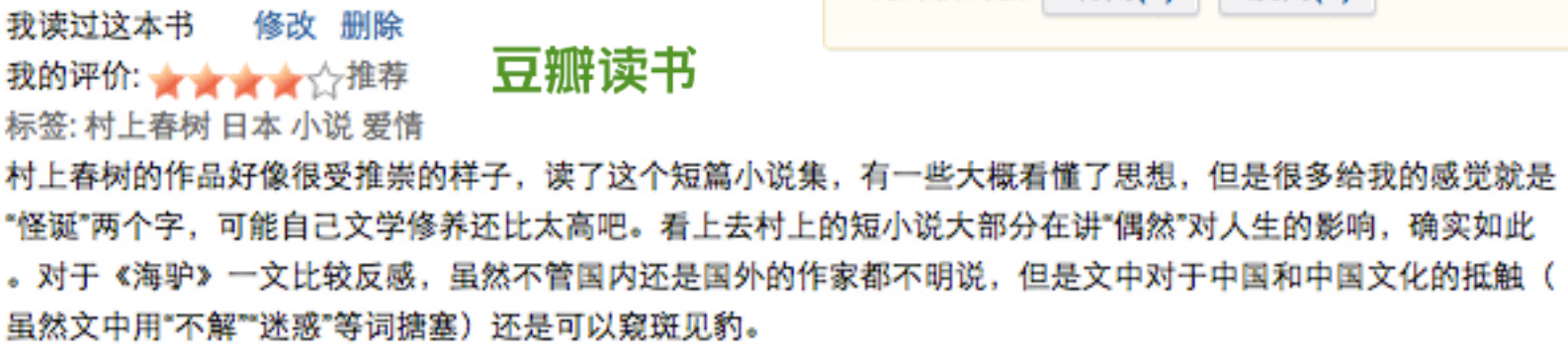
- User & User Profile(cont.)
  - 利用Item Profile构建User Profile
    - 出发点：计算User Profile和Item Profile 的相似度是常见操作
    - Case Study: 一种最简单的构建方法
      - 个性化电影推荐：

	Movie1	Movie2	Movie3	User Profile
类别	温情	温情	战争	(0.75温情, 0.25战争)
时长	90	120	90	100
打分	5	4	3	

    - 可以分别考察user在不同时间段内的profile，以反应user apatite的变化
      - 论文[2]正是基于这一思路
      - 以个性化查询推荐为背景
  - 优点：构建方便、易于使用(计算相似度)、巧妙地避开隐私信息、获得一些难以直观刻画的user profile信息

推荐系统的输入(cont.)

- Review(user 对 item 的评价)
  - 最简单的Review: 打分(Rating)
    - 一般是1~5的星级
  - 其它Review
    - 显式
      - 评论
      - 评分
      - 标签
    - 隐式
      - 查看历史记录
      - 购买记录
      - 页面停留时间



推荐系统的输出

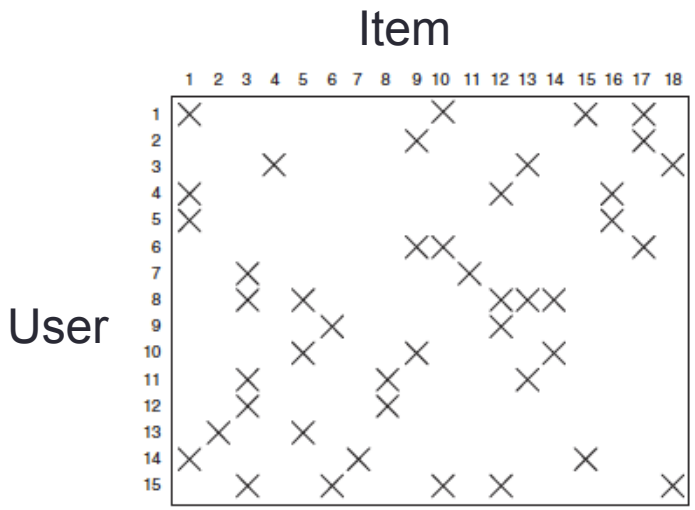
- 推荐列表(Recommendation List)
  - 按照特定的排序给出对该用户的推荐

- 推荐理由
  - 与 IR 系统的不同
  - 举例
    - e.g. 购买了某物品的用户有90%也购买了该物品
    - 该物品在某类别中人气最高
    - ……
  - 重要性
    - 解决推荐的合理性问题
    - 受到越来越多的重视[3]

- [3] A Survey of Explanations in Recommender Systems, IEEE Data Engineering. 2007.

推荐问题的形式化

- 基于User-Item Rating Matrix的形式化
  - 源自论文[1](GroupLens System)
  - A **sparse** matrix (user-item rating matrix)
    - 亚马逊书城：即使最活跃的用户，购书量也不到图书总量的1%
      - 1% * 2 million = 20 thousand
    - GroupLens实验数据集：
      - 只选择打分数>=20的user
      - 100,000 ratings / (934 users * 1682 items) = 6.3%
    - 越来越sparse
    - Rating
      - 0~5, 0 代表user未使用过该item
      - 1 最不喜欢
      - 5 最喜欢
  - Profiles



推荐系统的两大核心问题

- 执行推荐，两大核心问题
  - 预测(Prediction)
    - 预测每一个0处的可能值，即该 user 对该 item 可能的打分(越准越好)
    - 预测方法，评价方法
  - 推荐(Recommendation)
    - 依据Prediction环节的结果推荐用户未尝试过的item
    - 核心：Ranking
    - 策略众多
      - 直接按照Prediction给出的预测分值大小?
      - 用户的年龄段、历史爱好等Profile(User Profile一般在此派上用场)
      - 用户最近一段时间的购买记录……
    - 研究力度
      - 多数论文在给出 Prediction 的准确度后即停止
      - 推荐多样性[4]、推荐系统界面[5]

- [4] Improving Recommendation List Through Topic Diversification, WWW 2005
- [5] Is Seeing Believing? How Recommender Interfaces Affect Users' Opinions, CHI 2003

推荐方法分类及典型算法

- 推荐方法分类及典型算法
  - 多种分类方法
  - 典型算法介绍与对比

不同的分类方法

- 依据结果是否因人而异
  - 大众化推荐
    - 与用户本身及历史信息无关
    - 同样的外部条件下不同人获得同样的推荐
  - 个性化推荐
    - 推荐结果与用户本身的历史记录或行为有关

- 依据推荐算法的不同
  - 基于内容的推荐(Content-based Recommendation Algorithms)
  - 基于协同过滤的推荐(Collaborative Filtering-based)
  - 基于二部图的推荐(Structure-based)
  - 混合型推荐(Hybrid Recommendation Algorithms)

不同的分类方法(cont.)

- 细说基于协同过滤的推荐
  - 基于用户的协同过滤(User-based)
  - 基于物品的协同过滤(Item-based)
  - 基于社交网络关系的推荐(Social-based)
  - 基于模型的推荐(Model-based)
    - 基于矩阵分析(SVD/NMF, etc)
    - 基于机器学习(决策树、贝叶斯分类器、人工神经网络)
  - 基于关联规则的推荐(Association Rule Mining for Recommendation)

典型算法介绍:Content-based

- Content-based 推荐算法
  - Content? Profile!
  - 基于User-Profile & 基于Item-Profile

- 基于User-Profile
  - 又名Demographic-based(基于人口统计学的)[6]
  - 基本假设：**一个用户可能喜欢与其相似的用户所喜欢的物品**
  - 基本思想
    - 利用User Profile计算用户之间的相似度
    - 取出与该用户最相似的前K个用户
    - 将这K个用户的所覆盖的Item作为推荐列表
    - 以Item的平均得分为依据对列表进行排序

- [6] A framework for Collaborative, Content-Based and Demographic Filtering, AI Review, 1999

典型算法介绍:Content-based(cont.)

- 基于User Profile
  - 优点
    - 计算简单
      - User Profile相对固定，可实现线下计算，实时响应
  - 缺点
    - 可信度低
      - 性别、年龄等Profile相似的人完全可能有不同的偏好(准确)
      - 拥有相同偏好的人其Profile完全有可能很不同(召回)
    - 推荐结果可解释性不够
      - “与你具有相似属性的用户购买了 XX 商品”：难以让人信服
      - 给出相似人群的Profile？NO！隐私问题
  - 很少单独用来做推荐
    - 一般用于推荐结果的后期过滤

典型算法介绍:Content-based(cont.)

- 基于Item Profile
  - 基本假设：用户可能喜欢与他曾经喜欢过的东西相似的物品
  - 基本步骤(Simplified Version):
    - 利用Item Profile构建User Profile
      - 考虑该 User 所有打过分的 Item 的加权平均
      - 考虑User在不同时段打过分的Item的加权平均，线性拟合[2]
      - ……
    - 计算其它 Item 的 Item Profile 与该 User Profile 的相似度
    - 按照相似度大小给出推荐列表
  - 其它方式
    - 绕过计算 User Profile 的环节，直接使用 Item Profile
      - 以 Item 相似度作为权重，对该 User 的打分进行加权平均，计算预测打分
    - 转化为分类问题
      - 以(Item Profile, Like or not)作为训练数据，构建二分类器
      - 以(Item, Rating)作为训练数据，构建多分类器

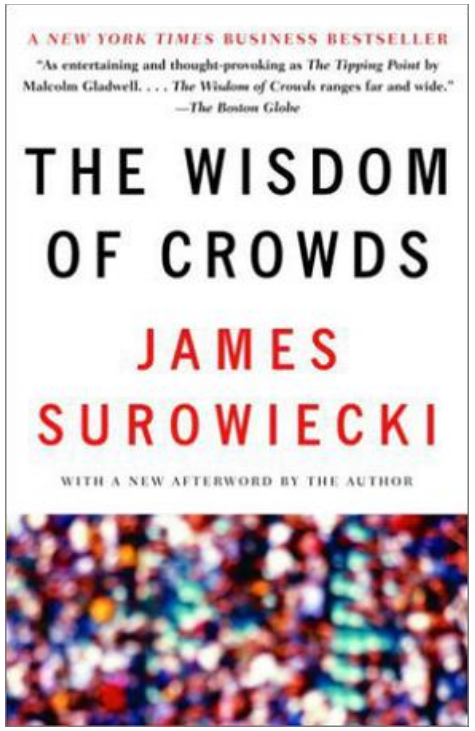
- [2] Adaptive Web Search based on User Profile Constructed without Any Effort from Users, WWW2004.

典型算法介绍:Content-based(cont.)

- 基于Item Profile
  - 优点
    - 对于新加入的Item没有冷启动(Cold Start)的问题
      - What is Cold-Start? ->
      - 只要有 Item Profile 就可以计算相似度
    - 推荐结果具有较好的可解释性
      - 该物品与你之前喜欢的某物品相似，如何相似
  - 缺点
    - 需要复杂模块甚至手工处理 Item Profile
      - 一系列复杂的算法 - 各种Tags相关的数据挖掘问题(Social Tags, Annotation 问题...)
      - 大量的人力投入，甚至领域专家的参与
        - Pandora: 音乐基因工程 - 每月标注10000首歌曲，每首需20min，已经标了10年。
    - 无法推荐用户不熟悉具有潜在兴趣的物品
      - 新颖性
    - 可扩展性不足
      - 不同的领域 Profile 几乎完全不同

典型算法介绍: CF-based

- 协同过滤(Collaborative Filtering)
  - 群体的智慧！ (Wisdom of the Crowd)
- Good or Bad?
  - Perhaps
    - 发现**或**使用我们难以描述的规律
    - Volinsky : 它可能找到我们从未意识到或为其命名的美观性，但从数学意义上讲，必须承认它是存在的
  - Or Perhaps Not
    - 群体结论有时会不靠谱
      - 2006年，沃尔玛的推荐引擎将《人猿星球》和《马丁·路德·金》纪录片捆绑推荐，为此而受到种族歧视的指控
    - Volinsky : 尝试预测人类行为不可避免地会出现一些错误



典型算法介绍: CF-based(cont.)

- User-based Recommendation
  - 实际上就是GroupLens[1]中提出的方法
  - 基本假设:
    - 一个用户可能喜欢与其具有相似爱好的用户所喜欢的物品
      - Content-based by User Profile (Demographic based)
    - What is “相似爱好”
    - 利用用户的打分历史记录计算用户相似度(行向量)
      - 具有相似偏好的用户，其在Item上的打分情况倾向于更相似
  - 基本步骤:
    - 设想对某一个user进行推荐
    - 数据预处理
      - Normalization, etc.
    - 计算它的Top K relevance Users(行向量)
      - 向量距离、夹角余弦相似度、Pearson相关性系数、etc.

典型算法介绍: CF-based(cont.)

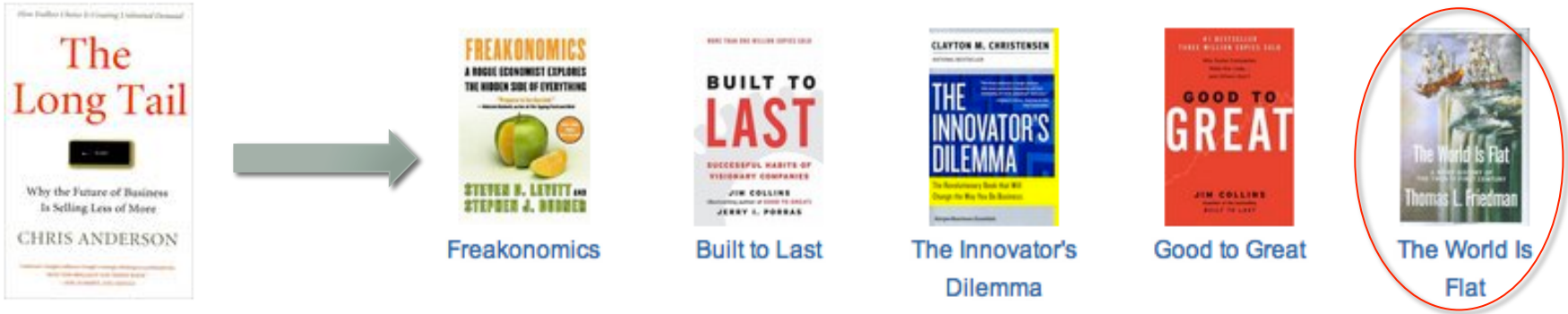
- User-based Recommendation
  - 基本步骤(cont.)
    - Top-N推荐
      - 统计这前 K 个用户中，出现频率最高且目标用户未体验过的 Item
      - 构建推荐列表
      - 推荐社区内的热门物品
    - 关联推荐
      - 将前 K 个用户购买的物品看做 K 个集合
      - 给定支持度和置信度，进行关联规则挖掘
      - 得到关联规则
        - If A, B, C then D
      - 依据目标用户的记录执行推荐
      - 购买了X的用户还购买了X

典型算法介绍: CF-based(cont.)

- User-based Recommendation
  - 优点
    - 避开了对 Profile 的挖掘
    - ……
  - 缺点
    - 用户数 >> 商品数
      - Top-K relevance user 的计算很耗时
        - 且难以线下计算
        - 且 user 数量不断增加
    - Cold Start 问题
      - 新用户加入时，几乎没有打分，难以个性化推荐
        - Content based 的优势体现出来
    - 不善于发现长尾
      - 总是倾向于推荐热门的 item

典型算法介绍: CF-based(cont.)

- Item-based Recommendation[7]
  - 亚马逊的专利算法
  - 亚马逊网络商城推荐系统的底层核心算法
  - 与 User-based 方法有某种“对称性”
    - 首次把GroupLens的方法称为User-based方法
  - 出发点:
    - 试图解决user-based方法中用户数巨大，计算top-k relevance耗时的问题
  - 基本假设:
    - 用户可能喜欢和他之前喜欢的物品相似的物品



- [7] Sarwar, Karypis, etc. Item-based Collaborative Filtering Recommendation Algorithms, WWW 2001

典型算法介绍: CF-based(cont.)

- Item-based Recommendation(cont.)
  - 基本步骤
    - 1. 利用user-item rating matrix，计算item之间的相似度(列向量)
      - 夹角余弦相似度、Pearson相关性系数、etc.
      - 只考虑共现的打分

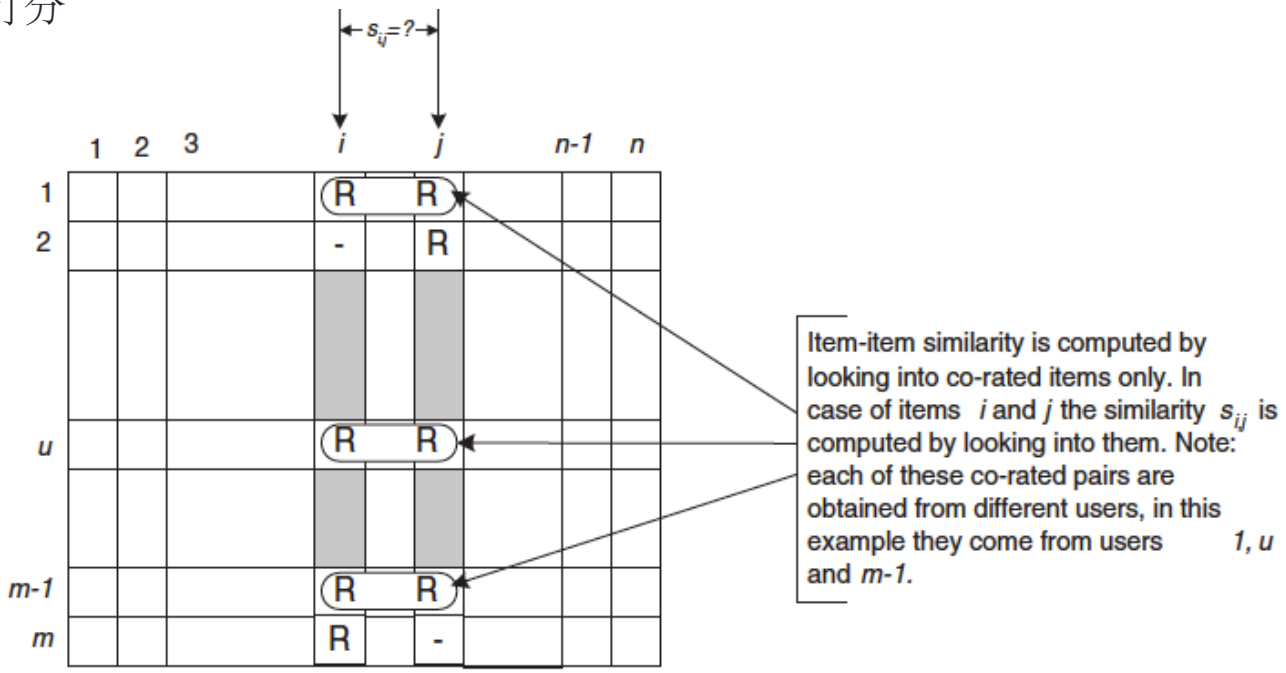
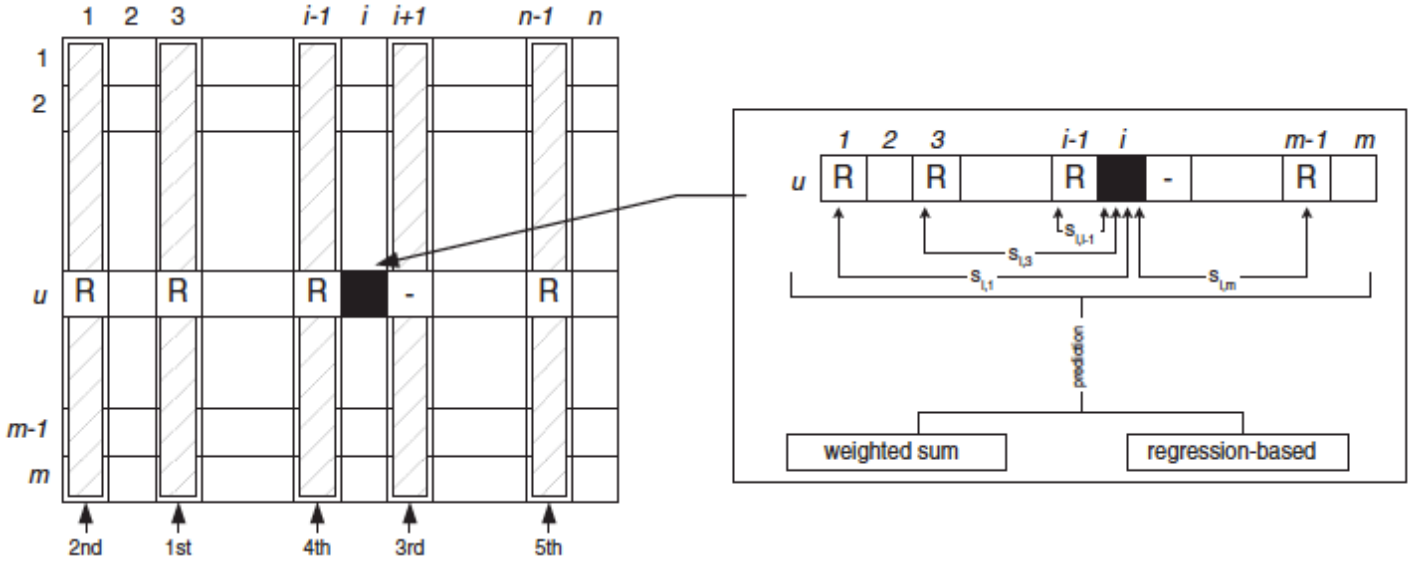


Figure 2: Isolation of the co-rated items and similarity computation

典型算法介绍: CF-based(cont.)

- Item-based Recommendation(cont.)
  - 基本步骤
    - 2.进行预测
      - 加权和预测（Weighted Sum）
        - 以相似度为权重，对用户打过的分加权平均
      - 线性回归预测（Regression Model）
        - 直接算相似度不准，先用$\bar{R'_N} = \alpha \bar{R}_i + \beta + \epsilon$ 做线性回归近似，再算相似度，再加权平均



典型算法介绍: CF-based(cont.)

- Item-based Recommendation(cont.)
  - 优点
    - 预测精度较 User-based 方法略高
      - MAE：Item 0.735 v.s. User 0.741
    - 可线下完成
      - Item变化剧烈程度远低于User，故Item相似度计算可线下完成，定期更新
    - 可实时响应
      - 用户添加新商品后，可立即给出新的推荐
    - 可解释性较好
      - 用户总是了解自己的购物历史
      - 给出被推荐物品的 Item Profile 没有隐私的问题
  - 缺点
    - 对 New Item 有Cold-Start的问题[8]
      - 从用户体验的角度，比 New User 的 Cold-Start 问题要好一些
- [8] Functional Matrix Factorizations for Cold-Start Recommendation, SIGIR 2011.

典型算法介绍: CF-based(cont.)

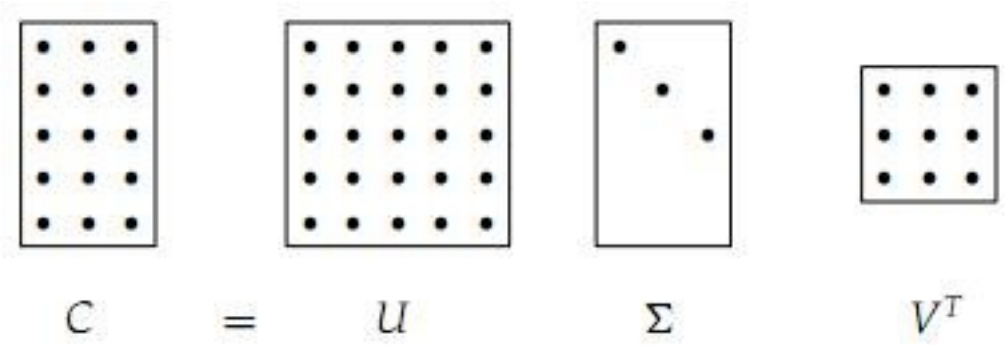
- User-based v.s. Item-based
  - 在线计算量
    - User 数 >> Item 数，且 Item 数据相对比较稳定 -- Item-based
      - 网络购物平台
    - Item 数 >> User 数，且 Item 数据更新频繁 -- User-based
      - 新闻、博客、微博推荐等
    - 两篇代表论文提出的背景
      - 1994 GroupLens 论文：网络新闻推荐
      - 2001 Item-based 论文：亚马逊网络商城
  - 应用场景
    - 非社交型网络： Item-based
      - 解释为 “和你具有相似兴趣的人也喜欢” v.s. “和你之前喜欢的某物品相似”
    - 社交型网络： User-based
      - 基于 User Network 和 Influence Network 的 User-based 方法，令用户信服

典型算法介绍: CF-based(cont.)

- User-based v.s. Item-based(cont.)
  - 推荐多样性
    - 互补性
      - 分别利用 User-based 和 Item-based 得到推荐列表
      - 约50%相同，50%不同，但却具有相似的精度 – 互补
    - 多样性
      - 用户多样性：单看一个user得到的推荐列表中item的两两相似度
        - User-based 好于 Item-based : Item-based 倾向于推荐和过去相似的物品
      - 系统多样性(即整个系统被推荐到的 item 的多样性)
        - Item-based 好于 User-based : User-based 倾向于推荐热门的物品
  - 用户对推荐算法的适应度
    - User-based
      - 用户适应度 正比于 与其有共同喜好的用户数量，即“大众性”
    - Item-based
      - 用户适应度 正比于 其自身所喜好的物品的自相似度，即“一致性”
  - Combination!

典型算法介绍: CF-based(cont.)

- Model-based Recommendation
  - 矩阵分解的道路
    - SVD低秩逼近



$$||X||_F = \sqrt{\sum_{i=1}^{M}\sum_{j=1}^{N} X_{ij}^2}.$$

1. 给定一个矩阵C，对其奇异值分解：$C = U\Sigma V^T$

2. 构造Σ_k，它是将Σ的第k+1行至M行设为零，也就是把Σ的最小的r-k个(the r-k smallest)奇异值设为零。

3. 计算C_k：$C_k = U\Sigma_k V^T$

典型算法介绍: CF-based(cont.)

- Model-based Recommendation(cont.)
  - 矩阵分解的道路(cont.)
    - SVD分解的结果中存在负值!
    - NMF(Non-negative Matrix Factorization)[9]
      - 停止条件：非零值处的方差小于指定阈值(本例<0.001)

$$R \approx P \times Q^T = \hat{R}$$

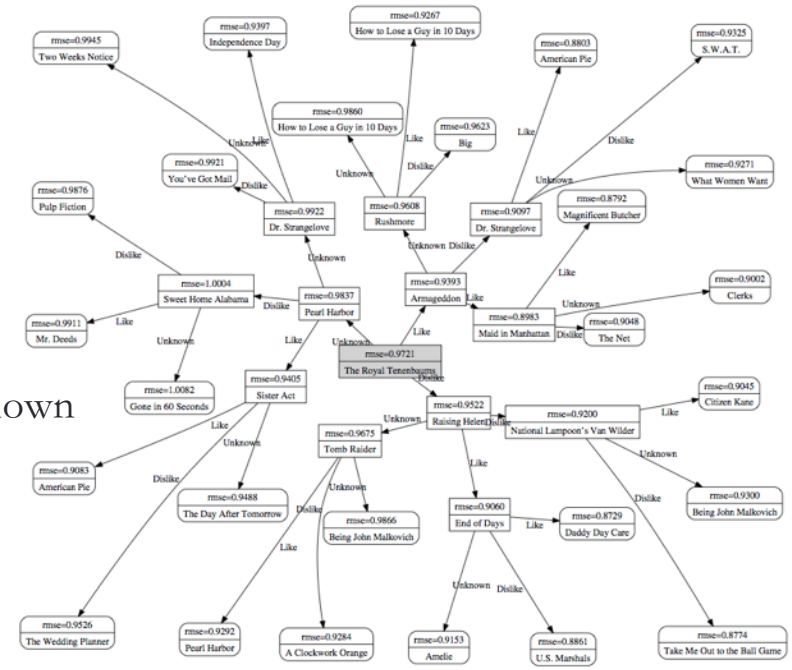
	I1	I2	I3	I4
U1	5	3	-	1
U2	4	-	-	1
U3	1	1	-	5
U4	1	-	-	4
U5	-	1	5	4

	I1	I2	I3	I4
U1	4.97	2.98	2.18	0.98
U2	3.97	2.40	1.97	0.99
U3	1.02	0.93	5.32	4.93
U4	1.00	0.85	4.59	3.93
U5	1.36	1.07	4.89	4.12

- [9] Algorithms for Non-negative Matrix Factorization. MIT Press, 2001.

典型算法介绍: CF-based(cont.)

- Model-based Recommendation
  - 机器学习的道路
    - 利用用户评分数据训练决策树[10]
      - 建树
        - 用户打分 -> dislike(1~3) like(4~5) unknown
        - 构建决策树
        - 终止条件
          - 达到设定的最大深度
          - 当前节点的最少user评分数量



      - 预测
        - 对于某个user，按照其打分情况将其一步步映射到某个叶子节点(user集合)
        - 用该user集合中对个item打分的均值作为预测
          - 论文进一步讨论了更精细的预测方法以解决过拟合的问题

- [10]Adapting Boost strapping of Recommender System using Decision Tree, WSDM 2011.

典型算法介绍: CF-based(cont.)

- Model-based Recommendation
  - 优点
    - 响应迅速
      - 只要训练出模型，即可快速判断
    - 缺点
      - 增量训练问题
        - 如何把用户新增或实时的喜好反馈给模型
          - 增量式SVD
          - 决策树的增量学习

典型算法介绍: Hybrid

- Hybrid Recommendation(混合型推荐系统)
  - 大部分商业推荐系统都是混合型的
  - Netflix大奖赛BellKor Pragmatic Chaos队 & The Ensemble 队
    - 100多个协同过滤算法的融合

  - 小八卦
    - 中科院自动化所
      - The Ensemble 队
      - 项亮
      - 公开测试集第一
      - 隐藏测试集第二
      - 整体排名第二

典型算法介绍: Hybrid(cont.)

- Hybrid Recommendation(cont.)
  - 加权融合(Weighted Merge)
  - 切换(Switch)
  - 混合(Mix)
  - 级联(Cascade)
  - 特征组合(Feature Combination)
  - etc

推荐系统的评价指标

- 评价方法
- 评价指标

推荐系统的评价指标(cont.)

- 评价方法
  - 离线评测
  - 在线评测
  - 用户调研

- 与IR评测有类似之处
  - 但Hulu的指出[11] [注]
    - Position Bias 假设在推荐系统中并不完全适用
      - 用户并不会因为某个Item在推荐列表里排第一就去点击
    - NDCG等指标不完全适用
      - Hulu的实验:在以NDCG为指标的线下评测中表现好的算法，线上点击率未必高
      - 结论：把点击率作为在线评估指标要谨慎(???)

- [11] Do Clicks Measure Recommendation Relevancy? An Empirical User Study. RecSys 2010.
- [注] http://hi.baidu.com/chen_1st/blog/item/2ec003628d788ce2f736541a.html (by chen_1st ?)

推荐系统的评价指标(cont.)

- 常见评估指标
  - 准确性(accuracy)
    - 预测准确度
      - MAE / NMAE
      - RMSE / ARMSE
      - MAP
    - 决策支持准确度
      - 相关度Correlation
        - Pearson / Spearman / Kendall Tau
      - Reversal Rate
      - Precision/Recall/F-measure
      - ROC曲线
  - 可用性(usefulness)
    - Diversity
      - Shannon Entropy / Gini Index
- 其它
  - 新颖性
  - 鲁棒性
  - 自适应性
  - 可扩展性
  - 推荐效率
  - 可解释性
  - ……

推荐系统的评价指标(cont.)

- MAE(Mean Absolute Error 平均绝对误差)

$$\mathrm{MAE} = \sqrt{\frac{1}{|\mathscr{T}|}\sum_{(u,i)\in\mathscr{T}}|\hat{r}_{ui} - r_{ui}|}$$

  - 直观解释
    - Step1：准备Data Set，构建新矩阵
      - Random(80%显示，20%隐藏，用于被预测)
      - Simulink(如果有时间信息的话)
    - Step2：利用预测算法预测新矩阵的空白值
    - Step3：利用20%被隐藏数值的真实值和预测值进行评价

推荐系统的评价指标(cont.)

- RMSE(Root Mean Square Error, 根/均/方差)

$$\mathrm{RMSE} = \sqrt{\frac{1}{|\mathscr{T}|}\sum_{(u,i)\in\mathscr{T}}(\hat{r}_{ui} - r_{ui})^2}$$

  - 目前学术论文中应用最广泛
  - Netflix大奖赛的最终评价指标(10% promotion -> $1 million)
  - 与SVD分解的数学关系
    - SVD低秩逼近就是要找一个秩为k(<r)且与原矩阵的差值矩阵F范数最小的近似矩阵
    - 稍有不同

$$||X||_F = \sqrt{\sum_{i=1}^{M}\sum_{j=1}^{N} X_{ij}^2}.$$

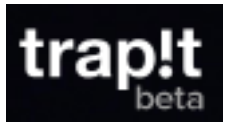
典型商业推荐系统

- 典型商业推荐系统
  - 典型商业推荐系统

典型商业推荐系统浅析(cont.)

- 电子商务领域

- 新闻与阅读

- 电影视频

- 音乐

面临的问题及发展方向

- 面临的问题及发展方向[12]
  - 推荐系统目前面临的问题
  - 推荐系统潜在的发展方向

- [12] Recent Advances in Personal Recommender Systems. Inter. Jour. of Inform. And System Sci. 2009.

面临的问题及发展方向(cont.)

- 数据稀疏性(Data Sparsity)
  - 表现
    - Cold-Start 冷启动问题
      - New-User / New-Item
    - Neighbor Transitivity Problem（近邻传递问题？？）
      - 一些Item之间由于共同打分很少，故难以计算相似度
    - Long tail
      - 如何识别小众需求并进行推荐？一些Item打分本身就很少
  - 一些应对方法
    - 降维技术
      - SVD / PCA …
      - 但是不利于长尾推荐
    - 使用Hybrid推荐模型
      - 推荐方法之间互相弥补（e.g. User-based + Item-based）

面临的问题及发展方向(cont.)

- 同义词问题(Synonymy)
  - 表现
    - 同样的物品有不同的名字
      - 进一步导致了数据稀疏性
  - 一些应对方法
    - 基于统计/语法的同义词挖掘
    - LSI 潜在语义分析

- Gray Sheep问题
  - 表现
    - 有的人偏好与其它任何人都不同
    - 注：Black Sheep – 偏好与一般人完全相反 - acceptable failure
  - 应对
    - Content-based + CF 混合推荐

面临的问题及发展方向(cont.)

- Shilling Attack(Anti Spam)
  - 表现
    - 有人故意给自己的物品打高分、给对手的物品打低分
  - 应对
    - 被动：采用 Item-based 方法，作弊者相对是少数
    - 主动：采用各种Anti Spam的方法……

面临的问题及发展方向(cont.)

- 潜在的发展方向
  - 基于User Review的推荐
    - 受限于文本挖掘、情感分析技术的发展
    - 用户评论数据较难获取
    - 2007年才开始有相关研究出现[13]
  - 推荐的可解释性
    - 相比 IR， 可解释性在推荐中尤其重要
  - 用户交互方式
    - 推荐列表?
  - 长尾推荐
    - 满足小众需求， 发掘用户潜在兴趣 - 带来更多收益

- [13] Informed Recommender: Basing Recommendations on Consumer Product Reviews, RecSys 2007

DEMO

- Demo
  - 数据集介绍
  - demo

DEMO

- Demo
  - 数据集介绍
    - 大众点评 30w 餐厅数据 + ~200w user comment

DEMO

- 基于Map Reduce的分布式网络爬虫
  - Hadoop MapReduce 分布式计算框架
  - Hadoop hbase 数据库
  - 适合于抓取“Item + 评论页” 的网站
    - 结构可扩展

Outline

- 推荐系统概述
  - 什么是推荐(个性化推荐)
  - 个性化推荐的发展历史
  - 个性化推荐问题的输入输出
  - 个性化问题的形式化
  - 推荐系统的两大核心问题
- 个性化推荐方法分类及典型算法
  - 多种分类方法
  - 典型算法介绍与对比
- 推荐算法的评价方法和指标
  - 评价方法
  - 评价指标
- 典型商业推荐系统浅析
  - 典型商业推荐系统
  - 亚马逊购物推荐浅析
- 面临的问题及发展方向
  - 推荐系统目前面临的问题
  - 推荐系统潜在的发展方向
- Demo
  - 数据集介绍
  - demo