CSDN
不 止 于 代 码

5.1 推荐系统简介

CSDN学院

2017年11月

大纲

- 推荐系统出现的背景
- 基于协同过滤的推荐
- 基于内容的推荐
- 推荐系统的评价
- 案例分析

推荐系统出现的背景

- 随着信息技术和网络技术的发展，人们已经从信息匮乏时代进入了信息过载时代
  - 每天有10w左右的新闻报道
  - 每秒钟优酷土豆爱奇艺搜狐腾讯B站会多出时长几百小时的视频
  - 淘宝京东亚马逊当当一天上架上百w商品
- 信息消费者和信息生产者都面临挑战:
  - 无明确需求
  - 信息过载

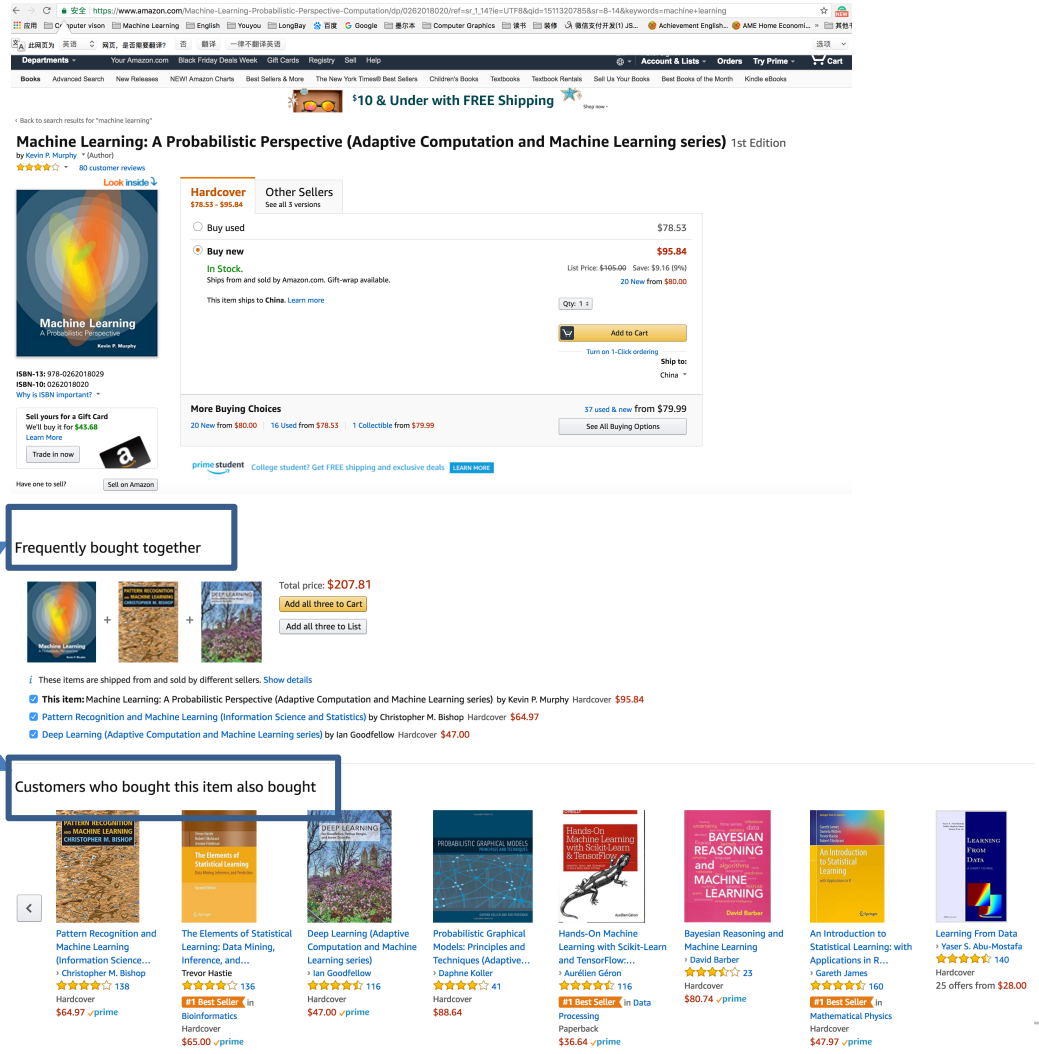
信息过载的解决方案

- 分类导航（1994年Yahoo…）
- 搜索引擎（ 1998年Google、2000年百度）
- 推荐系统（1994年明尼苏达大学GroupLens）

推荐系统

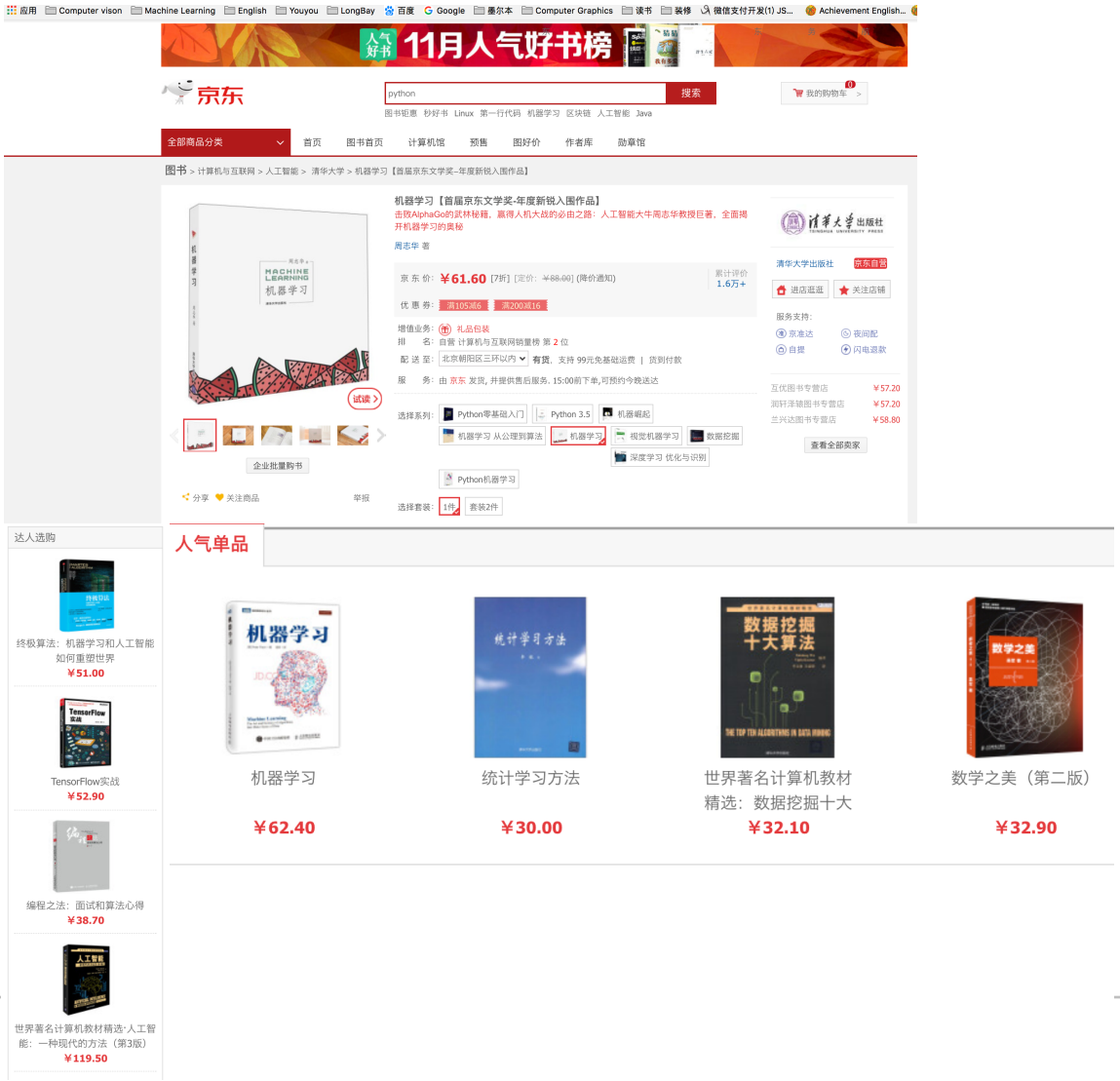
- 推荐系统在用户(User)和物品(Item)之间架设一座桥梁
  - 帮助用户发现对自己有价值的信息/物品
  - 让物品展现在对它感兴趣的用户面前

Amazon的推荐系统



推荐理由

京东的推荐系统


11月人气好书榜
京东
python
搜索
全部商品分类
首页
图书首页
计算机馆
预售
图好价
作者库
勋章馆
机器学习【首届京东文学奖-年度新锐入围作品】
京东价：￥61.60
达人选购
人气单品
机器学习
￥62.40
统计学习方法
￥30.00
世界著名计算机教材精选：数据挖掘十大
￥32.10
数学之美（第二版）
￥32.90
终极算法：机器学习和人工智能如何重塑世界
￥51.00
TensorFlow实战
￥52.90
编程之法：面试和算法心得
￥38.70
世界著名计算机教材精选·人工智能：一种现代的方法（第3版）
￥119.50

推荐系统 vs. 搜索系统

CSDN 不止于代码

- 相同点：提供帮助用户快速发现有用信息的技术
- 不同点
  - 搜索系统需用户主动提供关键词
  - 推荐系统无需用户提供明确需求，而是通过用户的历史行为给用户的兴趣建模
- 从某种意义上而言，推荐系统和搜索系统是互补
  - 搜索系统满足用户有明确目的的主动查找需求
  - 推荐系统能够在用户没有明确目的是帮助用户发现感兴趣的新内容

为什么使用推荐系统?

- 对用户
  - 发现感兴趣的东西 – 缩小选择范围
  - 帮助探索选择空间 – 发现新事
  - …
- 对商家
  - 对客户提供增值或独特的个性化服务
  - 增加用户的信任度和忠诚度 – 增加销售、点击率
  - 得到客户的更多信息
  - …

商家从推荐系统获利

- Netflix每年2/3的观看电影源自推荐
- Google news推荐系统能带来额外38%的点击
- 亚马逊每年35%的销售额都来源于推荐
- 头条半数以上新闻和广告点击来源于推荐
- 京东一年推荐和广告带来几亿的营收

NetFlix

- 2006年启动了Netflix大奖赛，旨在解决电影评分预测问题
  - 48w用户的1亿+打分和评论，17770部电影
- 推荐系统领域标志性事件:2009年，准确率提升10%，100w美金奖金
  - 2700+队伍

当一个会员访问Netflix，Netflix希望能够帮助他在几秒钟之内就找到他感兴趣的影片，以免他去寻找别的乐子。

https://www.douban.com/note/247017189/

Netflix Prize

The Nextflix prize story

- In October 2006, Netflix announced it would give a $1 million to whoever created a movie-recommending algorithm 10% better than its own.
- Within two weeks, the DVD rental company had received 169 submissions, including three that were slightly superior to Cinematch, Netflix's recommendation software
- After a month, more than a thousand programs had been entered, and the top scorers were almost halfway to the goal
- But what started out looking simple suddenly got hard. The rate of improvement began to slow. The same three or four teams clogged the top of the leader-board.
- Progress was almost imperceptible, and people began to say a 10 percent improvement might not be possible.
- Three years later, on 21st of September 2009, Netflix anno


Netflix Prize
Home
Rules
Leaderboard
Update
Congratulations!
The Netflix Prize sought to substantially improve the accuracy of predictions about how much someone is going to enjoy a movie based on their movie preferences.
On September 21, 2009 we awarded the $1M Grand Prize to team "BellKor's Pragmatic Chaos". Read about their algorithm, checkout team scores on the Leaderboard, and join the discussions on the Forum.
We applaud all the contributors to this quest, which improves our ability to connect people to the movies they love.

Leaderboard

Showing Test Score. Click here to show quiz score

Display top 20 **leaders.**

Rank	Team Name	Best Test Score	% Improvement	Best Submit Time
Grand Prize - RMSE = 0.8567 - Winning Team: BellKor's Pragmatic Chaos				
	BellKor's Pragmatic Chaos	0.8567	10.06	2009-07-26 18:18:28
	The Ensemble	0.8567	10.06	2009-07-26 18:38:22
	Grand Prize Team	0.8582	9.90	2009-07-10 21:24:40
	Opera Solutions and Vandelay United	0.8588	9.84	2009-07-10 01:12:31
	Vandelay Industries !	0.8591	9.81	2009-07-10 00:32:20
	PragmaticTheory	0.8594	9.77	2009-06-24 12:06:56
	BellKor in BigChaos	0.8601	9.70	2009-05-13 08:14:09
	Dace_	0.8612	9.59	2009-07-24 17:18:43
	Feeds2	0.8622	9.48	2009-07-12 13:11:51
0	BigChaos	0.8623	9.47	2009-04-07 12:33:59
11	Opera Solutions	0.8623	9.47	2009-07-24 00:34:07
12	BellKor	0.8624	9.46	2009-07-26 17:19:11
Progress Prize 2008 - RMSE = 0.8627 - Winning Team: BellKor in BigChaos				
13	xiangliang	0.8642	9.27	2009-07-15 14:53:22
14	Gravity	0.8643	9.26	2009-04-22 18:31:32
15	Ces	0.8651	9.18	2009-06-21 19:24:53
16	Invisible Ideas	0.8653	9.15	2009-07-15 15:53:04
17	Just a guy in a garage	0.8662	9.06	2009-05-24 10:02:54
18	J Dennis Su	0.8666	9.02	2009-03-07 17:16:17
19	Craig Carmichael	0.8666	9.02	2009-07-25 16:00:54
20	acmehill	0.8668	9.00	2009-03-21 16:20:50
Progress Prize 2007 - RMSE = 0.8723 - Winning Team: KorBell				
Cinematch score - RMSE = 0.9525				

阶段性竞赛结果

- 比赛开始一年后，Korbell 的团队以 8.43% 的提升赢得了第一个阶段奖。
- 其中两种最有效的算法
  - 矩阵分解（通常被叫做 SVD，奇异值分解）： RMSE为0.8914
  - 局限型玻尔兹曼机（RBM）：RMSE为0.8990
  - 两种方法线型融合：将RMSE 降低到 0.88
- 这两种算法应用到了产品中，而且作为推荐引擎的一部分一直被使用至今。

最终优胜方案

- 数百种预测模型被融合在一起，最终突破了 0.8572 的临界线。
- 但很遗憾，这些在比赛数据上胜出的算法，到了线上却表现不够出色。因此，并没有应用到我们的线上环境。

Netflix推荐系统的特性

- 为每个用户建立画像
- 推荐理由
- 多样性
- 好友推荐
- Genre系统
- 相似度系统

京东JData竞赛

- 高潜用户购买意向预测:根据历史数据(用户、商品和行为数据)，构建用户购买商品的预测模型，输出高潜用户和目标商品的匹配结果
  - 2017年3月-2017年6月
  - 7346人、四千多支团队参加了比赛

http://www.datafountain.cn/projects/jdata/

阿里移动推荐算法大赛

- 根据阿里巴巴移动电商平台的真实用户—商品行为数据和位置信息，构建面向移动电子商务的商品推荐模型，为移动用户在合适的时间、合适的地点精准推荐合适的内容
  - 2015年3月-2015年7月
  - 7186人参赛

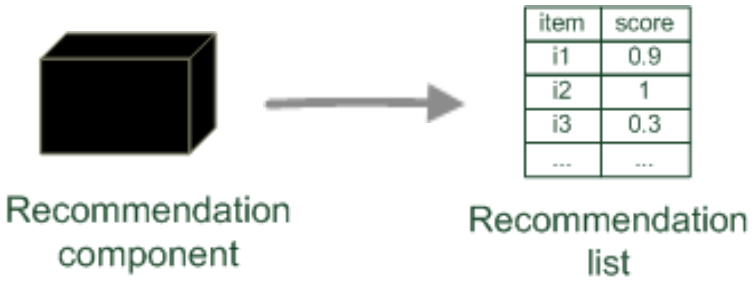
https://tianchi.aliyun.com/competition/information.htm?raceId=1

推荐系统定义

- 推荐系统可视为一个函数 $f: U \times S \rightarrow R$
- 输入：
  - 用户（User）模型（如兴趣、场景上下问、社交关系）
  - 物品（ Items）模型 (带有物品的描述信息)
  - 用户—物品之间的关系（如打分、浏览、点击、购买…）
- 计算：
  - 相关分数，用于排序
- 最后：
  - 推荐相关的物品
- 但是：
  - 请记住相关性可能是上下文相关的
  - 推荐列表的性质可能非常重要（如多样性）

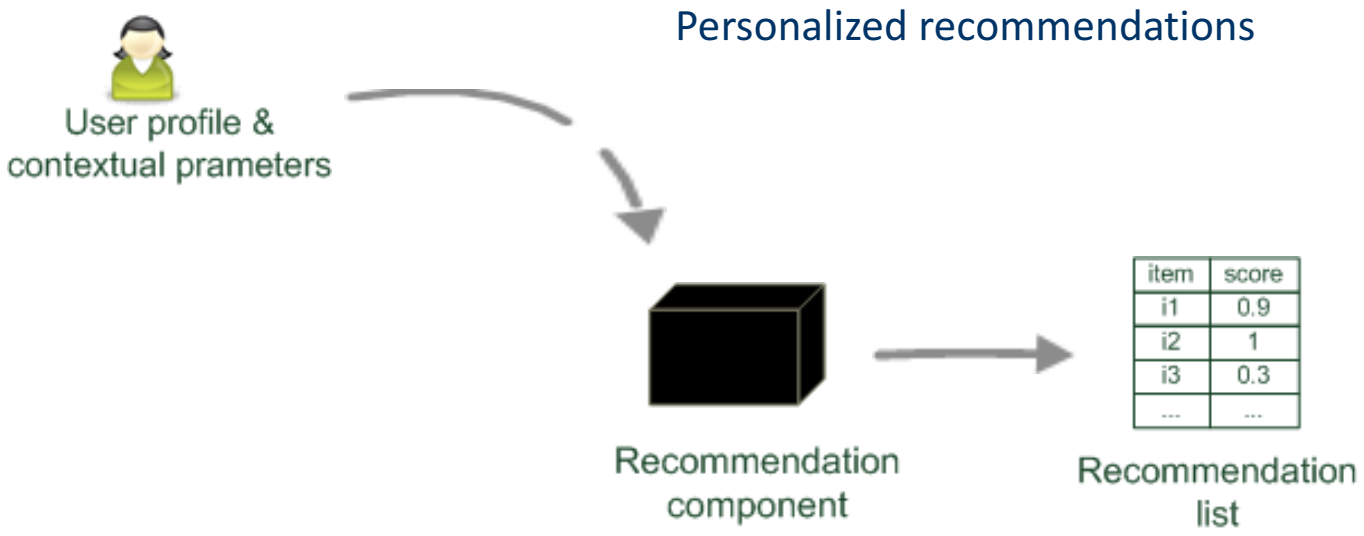
Paradigms of recommender systems

Recommender systems reduce information overload by estimating relevance

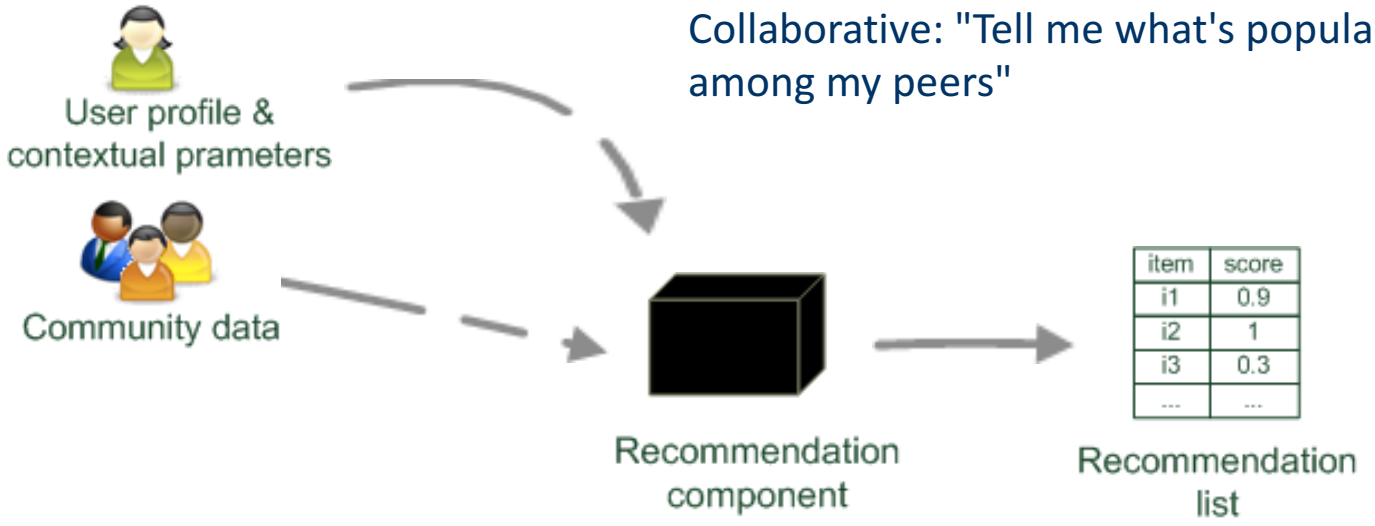


From： International Joint Conference on Artificial Intelligence 2013 Tutorial: Recommender Systems

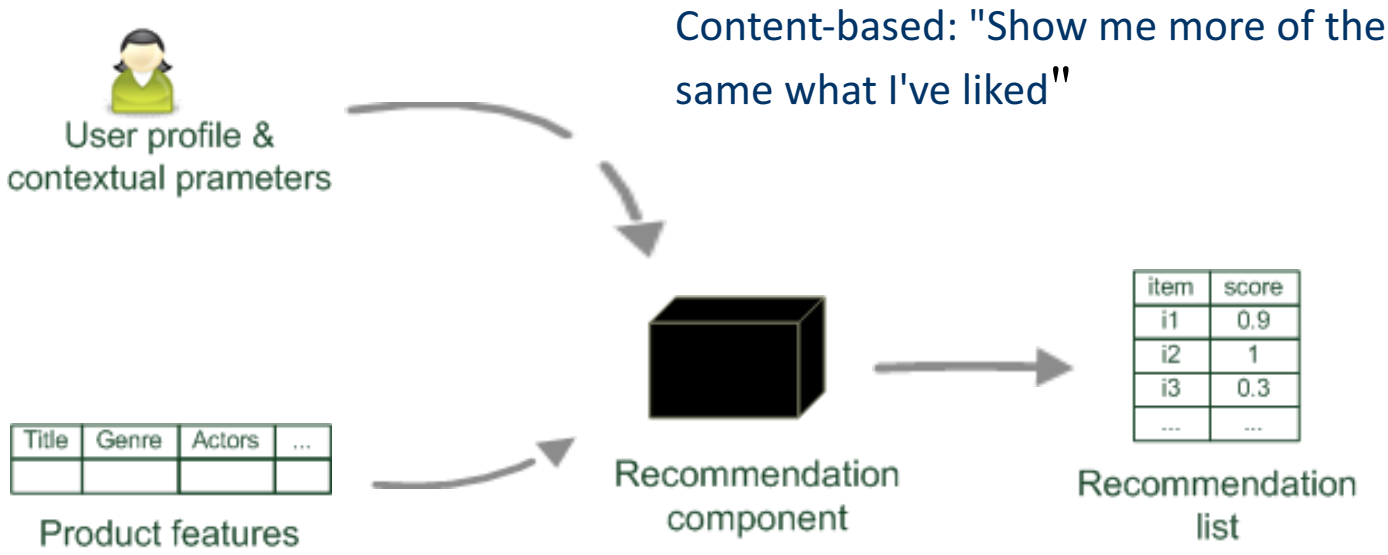
Paradigms of recommender systems



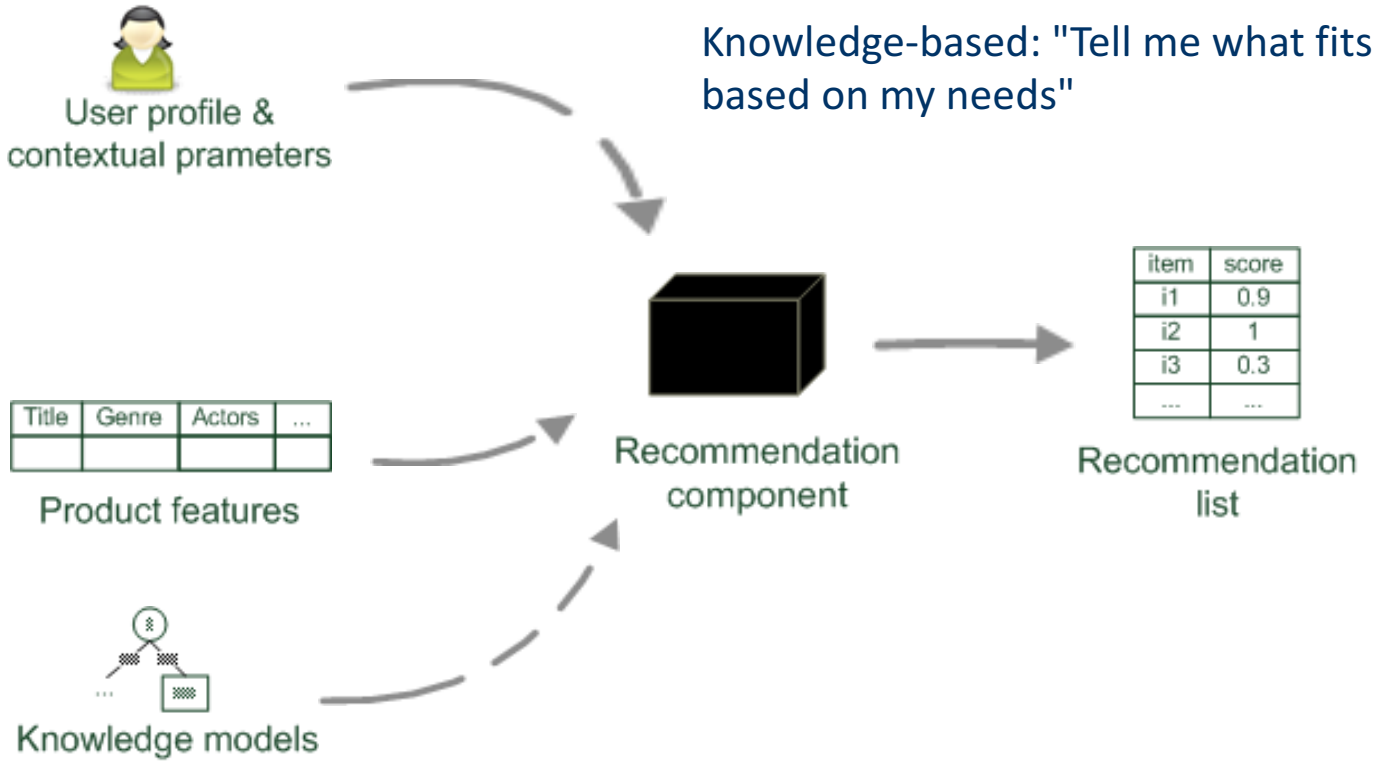
Paradigms of recommender systems



Paradigms of recommender systems

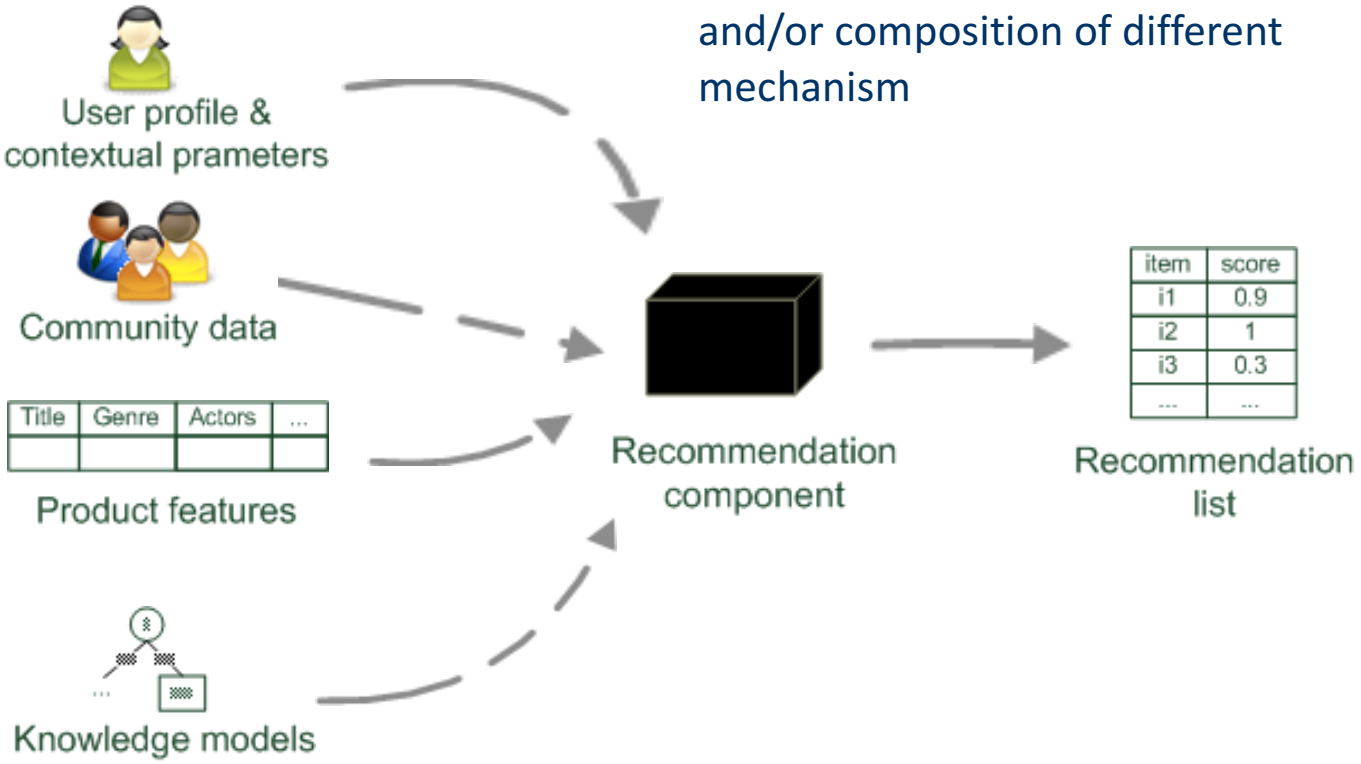


Paradigms of recommender systems



Paradigms of recommender systems

Hybrid: combinations of various inputs and/or composition of different mechanism



►例： Event Recommendation Engine Challenge

- https://www.kaggle.com/c/event-recommendation-engine-challenge/data

- 根据
  - events they've responded to in the past
  - user demographic information
  - what events they've seen and clicked on in our app
- 预测用户对某个事件是否感兴趣

数据

File Name	Available Formats
event_attendees.csv	.gz (55.03 mb)
events.csv	.gz (161.46 mb)
user_friends.csv	.gz (148.60 mb)
random_benchmark	.csv (146.91 kb)
users	.csv (2.63 mb)
event_popularity_benchmark	.csv (146.91 kb)
public_leaderboard_solution	.csv (6.46 kb)
test	.csv (574.40 kb)
train	.csv (924.23 kb)
event_popularity_benchmark_private_test_only	.csv (51.51 kb)

event_id,yes, maybe, invited, and no

event_id, user_id,start_time, city, state, zip, country, lat, and lng ， 101 columns of words count

user ， friends

user_id, locale, birthyear, gender, joinedAt, location, and timezone

user, event, invited, timestamp, interested, and not_interested

►例： Expedia Hotel Recommendations

- https://www.kaggle.com/c/expedia-hotel-recommendations

- 基于
  - 用户的搜索数据的一些属性
  - 酒店的属性（从以前的用户评论文本中提取）
- 预测他们会预定哪一个酒店

Train/Test文件数据字段

Column name	Description	Data type
date_time	Timestamp	string
site_name	ID of the Expedia point of sale (i.e. Expedia.com, Expedia.co.uk, Expedia.co.jp, ...)	int
posa_continent	ID of continent associated with site_name	int
user_location_country	The ID of the country the customer is located	int
user_location_region	The ID of the region the customer is located	int
user_location_city	The ID of the city the customer is located	int
orig_destination_distance	Physical distance between a hotel and a customer at the time of search. A null means the distance could not be calculated	double
user_id	ID of user	int
is_mobile	1 when a user connected from a mobile device, 0 otherwise	tinyint
is_package	1 if the click/booking was generated as a part of a package (i.e. combined with a flight), 0 otherwise	int
channel	ID of a marketing channel	int
srch_ci	Checkin date	string
srch_co	Checkout date	string
srch_adults_cnt	The number of adults specified in the hotel room	int
srch_children_cnt	The number of (extra occupancy) children specified in the hotel room	int
srch_rm_cnt	The number of hotel rooms specified in the search	int
srch_destination_id	ID of the destination where the hotel search was performed	int
srch_destination_type_id	Type of destination	int
hotel_continent	Hotel continent	int
hotel_country	Hotel country	int
hotel_market	Hotel market	int
is_booking	1 if a booking, 0 if a click	tinyint
cnt	Numer of similar events in the context of the same user session	bigint
hotel_cluster	ID of a hotel cluster	

用户基本信息

酒店基本信息

用户-酒店交互信息：搜索上下文

酒店信息描述文件destinations.csv

Column name	Description	Data type
srch_destination_id	ID of the destination where the hotel search was performed	int
d1-d149	latent description of search regions	double

Referneces

- https://wenku.baidu.com/view/0cd314cd0722192e4436f6aa.html
- http://www.xiaoxiansheng.com/html/talk/detail/3845?categoryId=319

CSDN
不 止 于 代 码

THANK YOU

AI100