



GOME

机器学习排序实践

邱宝军 杨骥 李冰青

国美在线大数据中心

2016年7月1日

国美大数据研究院

目录

➤ 01 机器学习排序

➤ 02 推荐排序实践



— 15 —



中国地质大学(Laurea) 张强 天津
WUJIAN (D)

推荐
vs.广告
vs.搜索



gitc



日期	照片	说明	备注	日期	照片	说明	备注
1993.10.1				1993.10.1			
1993.10.2				1993.10.2			
1993.10.3				1993.10.3			
1993.10.4				1993.10.4			
1993.10.5				1993.10.5			
1993.10.6				1993.10.6			
1993.10.7				1993.10.7			
1993.10.8				1993.10.8			
1993.10.9				1993.10.9			
1993.10.10				1993.10.10			
1993.10.11				1993.10.11			
1993.10.12				1993.10.12			
1993.10.13				1993.10.13			
1993.10.14				1993.10.14			
1993.10.15				1993.10.15			
1993.10.16				1993.10.16			
1993.10.17				1993.10.17			
1993.10.18				1993.10.18			
1993.10.19				1993.10.19			
1993.10.20				1993.10.20			
1993.10.21				1993.10.21			
1993.10.22				1993.10.22			
1993.10.23				1993.10.23			
1993.10.24				1993.10.24			
1993.10.25				1993.10.25			
1993.10.26				1993.10.26			
1993.10.27				1993.10.27			
1993.10.28				1993.10.28			
1993.10.29				1993.10.29			
1993.10.30				1993.10.30			
1993.10.31				1993.10.31			

[歷史](#)
[當前顯示方式](#)
[登錄](#)
[註冊](#)
[我的書架](#)
[我的書架](#)
[我的書架](#)

GTC全球互联网技术大会

What is G8TC? G8TC全球互联网技术大会以汇聚行业精英、促进技术交流、加强国际合作以及推动行业繁荣为大会宗旨。通过针对互联网技术领域有口无心的演讲并现场问答。

[大会日程](#) [关于大会](#) [注册购票](#) [嘉宾](#) [资料下载](#) [会场地图](#)

全球互联网技术大会 百度百科

2013年5月23日，由工业和信息化部、中国科学院、国家发展和改革委员会、人力资源和社会保障部联合启动的“互联网+”行动计划，旨在推动互联网与经济社会各领域的深度融合，促进经济转型升级。该计划强调通过互联网技术提升传统产业效率，培育新兴业态，并加强网络安全保障。相关报道可参见：http://www.gov.cn/jq/2013/05/23/content_2544917.htm。

Greater Illinois Title Company 翻译: 伊利诺伊大标题公司

[gitc 首页](#)、[文档和下载](#) - [代码托管平台](#) - [开源中国社区](#)

[相关标签](#)

© 2015

[illegible]

图 10-1-10 式

地址：新豐干

● 考友

◎ 正文

Q30 Q31

negative **gates**

摘要

404-673-0000 ext. 200

People You May Know

Big Data Analyst at TD

Connect

Attended The
Chinese University of
Hong Kong

Connect

Senior Quality Engineer at Glandore Systems

Connect

Graduate student at
Rutgers University

Connect

搜索GITC

GITC全球互联网技术大会

What is **GITC**? **GITC**全球互联网技术大会以汇集行业精英、促进技术交流、加深商务合作以及推动行业发展为大会宗旨。通过针对互联网技术领域专门设计的先进开放会议会 ...

www.thegitc.com ▼

全球互联网技术大会_百度百科

GITC（全球互联网技术大会）项目于2013年5月正式发起，通过针对互联网技术领域专门设计的先进开放会议会展内容及形式，联合iTech Club（中国互联网技术 ...

baike.baidu.com/view/10900771 2016-5-31

GITC 2014新闻发布会暨开票仪式在京召开|GITC_滚动新闻 ...

2014年8月6日“**GITC** 2014全球互联网技术大会新闻发布会暨开票仪式”于北京五洲皇冠国际酒店隆重召开。作为**GITC** 2014全球互联网技术大会筹备召开的正式启动 ...

tech.sina.com.cn > 新浪科技 > 滚动新闻 > 2014GITC... ▼ 2014-9-30

GITC (团体儿童智力测验)_百度百科

中文名 团体儿童智力测验 外文名 **GITC** 优点 客观、精练、易操作、经济快速等 主要特点 适用于团体快速施行,非语言量表

baike.baidu.com/subview/11415563/14249887.htm ▼ 2016-1-17

Greater Illinois Title Company 翻译此页



As the largest title agency in Illinois, **Greater Illinois Title Company** provides a single point of contact for all title and closing

Perfect

Excellent

Good

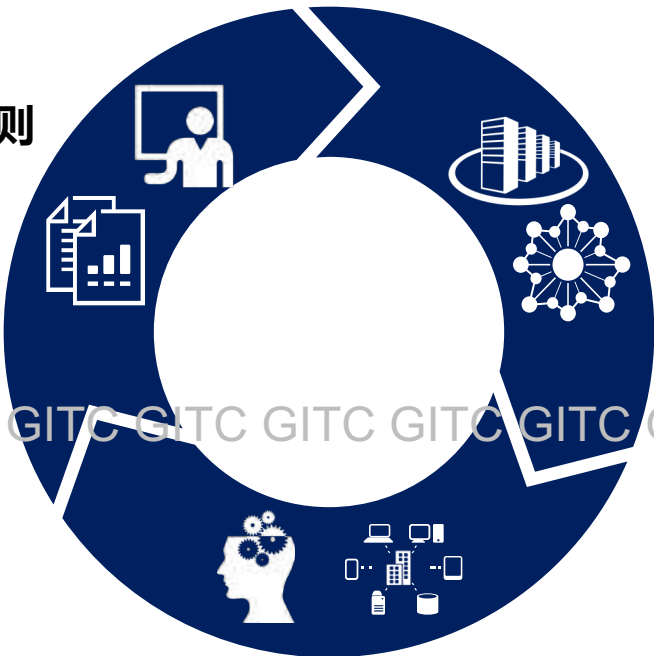
Fair

Bad

搜索词、目标文档(网站)、搜索人、时间、环境...

Query	URL	Label	Page-Rank	BM25	...	Feature_N
GITC	Thegitc.com	Perfect	0.6	0.9		100
GITC	...	Excellent	0.9	0.9		66
GITC	...	Good	0.8	0.7		75
GITC	...	Fair	0.9	0.6		32
GITC	...	Bad	0.2	0.65		120
国美	...	Perfect	0.6	0.8		150
国美	...	Perfect	0.5	0.7		50
...						

人工规则



人工规则 vs. 机器学习
(广告 vs. 搜索 vs. 推荐)

机器学习排序(Learning to Rank)

- ✓ Pointwise
排序问题转化为多类分类问题或者回归问题，相关度相同的为一类
- ✓ Pairwise
排序问题转化为二分类问题，由两两之间的偏序关系得到全局的排序
- ✓ Listwise
直接对排序结果进行优化

训练数据

<Query, Doc> → 标注

A	B	C	D	E	F	G
	Label			Label		
B	+		B-A	+		
E	+		B-C	+		
A	-		B-D	+		
C	-		C-E	-		
D	-		D-E	-		
			E-A	+		

特征工程

模型训练
(& 离线测试)

线上部署
(& A/B测试)



特征工程-从实体到特征

用户属性、
购买力、
品类倾向



用户

查询量、
转化率、
品类分



查询

UV、PV、
订单量、
转化率



商品

店铺评分、
店铺流量、
店铺评价



商家

当地温度、
污染指数、
人口数量



环境

.....

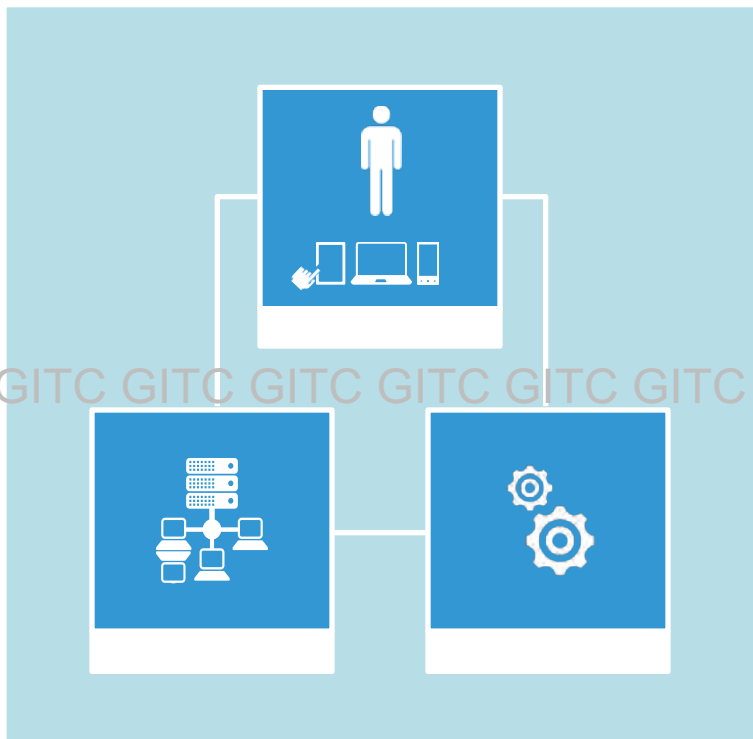
关键词匹配、价格匹配等

区域销量、商品天气匹配等

品类匹配、购买力匹配等

- 协同过滤
- 回归模型（LR,SVR...）
- 矩阵分解（SVD,LFM,SVD++,...）
- 马尔科夫链
- 聚类 (K-means, ...)
- Deep Learning（RBM, CNN, RNN, ...）
- LDA
- Association Rules
- GBDT/RF
- ...

离线预处理



在线处理



- 可量化目标：分类准确率、NDCG、点击率、转化率
- 假设→试验→结论



德鲁克：如果你不能衡量它，就无法增长它

国美推荐A/B测试配置及量化

第一步：更改A组权重



分组信息

分组名称: A组

权重: 10

分组配置策略

选择配置策略: 默认配置

确定 取消

第三步：完成分组流量配置及分组备注



分组列表

分组名称	当前权重	结束时间	开始时间	操作
A组	10	2005-04-20 18:53:42	2005-04-20 18:53:42	编辑
B组	20	2005-04-20 18:53:42	2005-04-20 18:53:42	编辑

添加 删除

第二步：更改B组配置及流量



分组信息

分组名称: B组

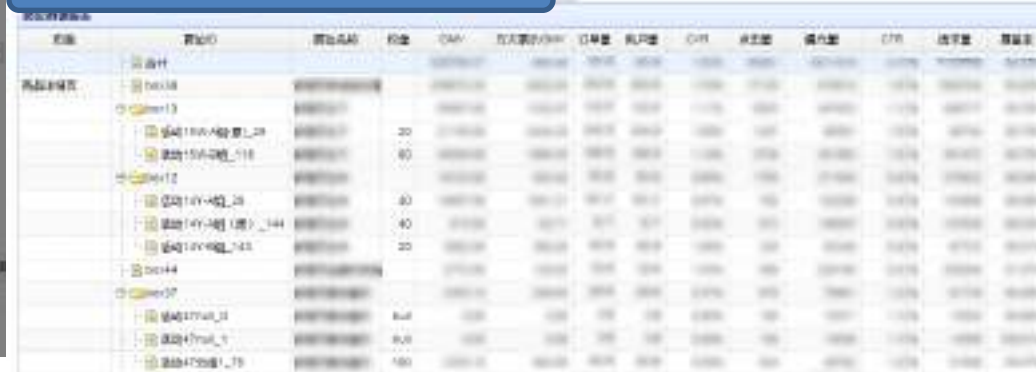
权重: 20

分组配置策略

选择配置策略: 默认配置

确定 取消

持续观察KPI指标



分组名称	KPI名称	KPI值	KPI单位
A组	转化率	1.2%	%
B组	转化率	1.5%	%
A组	点击率	2.1%	%
B组	点击率	2.3%	%
A组	停留时间	1.2s	s
B组	停留时间	1.5s	s

对比A/B测试效果

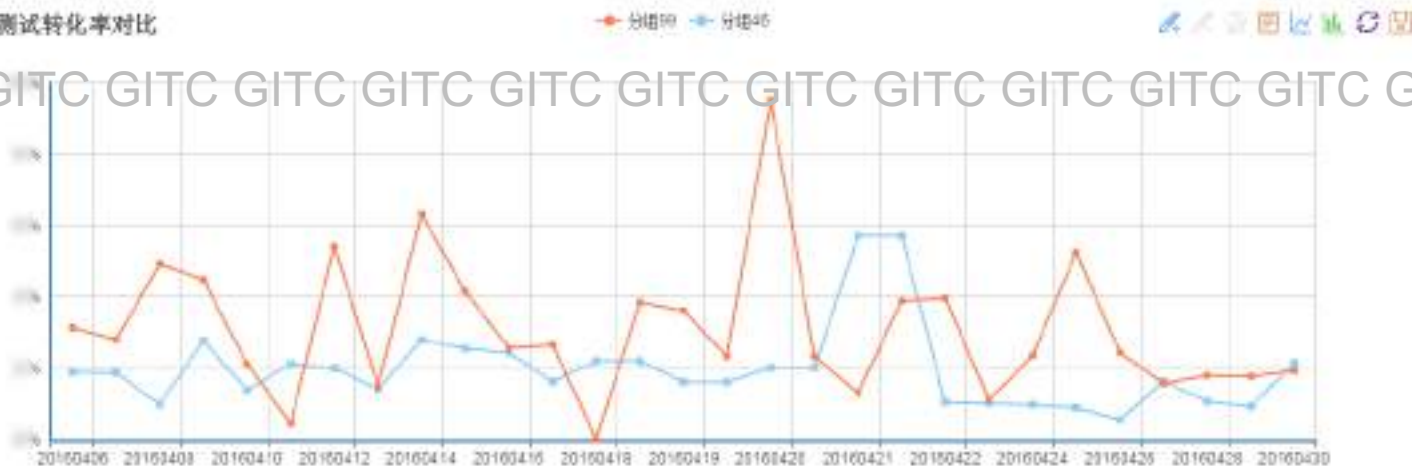
开始时间: 2016-04-06 结束时间: 2016-04-30 精确查询: 查询 对比

基础数据列表

展位编号	活动优先级	分组编号	分组权重	转化军位置	标准差	95%范围区间	数据开始日期	数据结束日期
10000	高	09	50	1	0.003	0.001-0.005	2016-04-06	2016-04-30
10000	高	46	50	1	0.003	0.001-0.005	2016-04-06	2016-04-30
10000	低	47	50	1	0.003	0.001-0.005	2016-04-06	2016-04-30

T检验的P值: 0.003 两组差异95%的范围: 0.001-0.005

A/B测试转化率对比



机器学习排序-小结

场景恢复

用户交互行为

用户特征

商品特征

环境特征

训练样本构建

样本标注

购买>购物车>关注>点击>展示

(购买力水平、长期偏好、性别、家庭属性、品类、品牌、产品词、修饰词、销量、点击量、转化率、浏览用户量、下单用户量等特征)

模型训练

数据清洗

离线评测
(正确率、召回率、NDCG)

特征工程
(特征预处理、特征选择、特征离散化)

参数调优

线上实验

用户实时行为

用户&商品画像

多模型召回商品

实时交互特征
长期交互特征

离线特征计算

在线特征计算

排序

A/B测试

目录

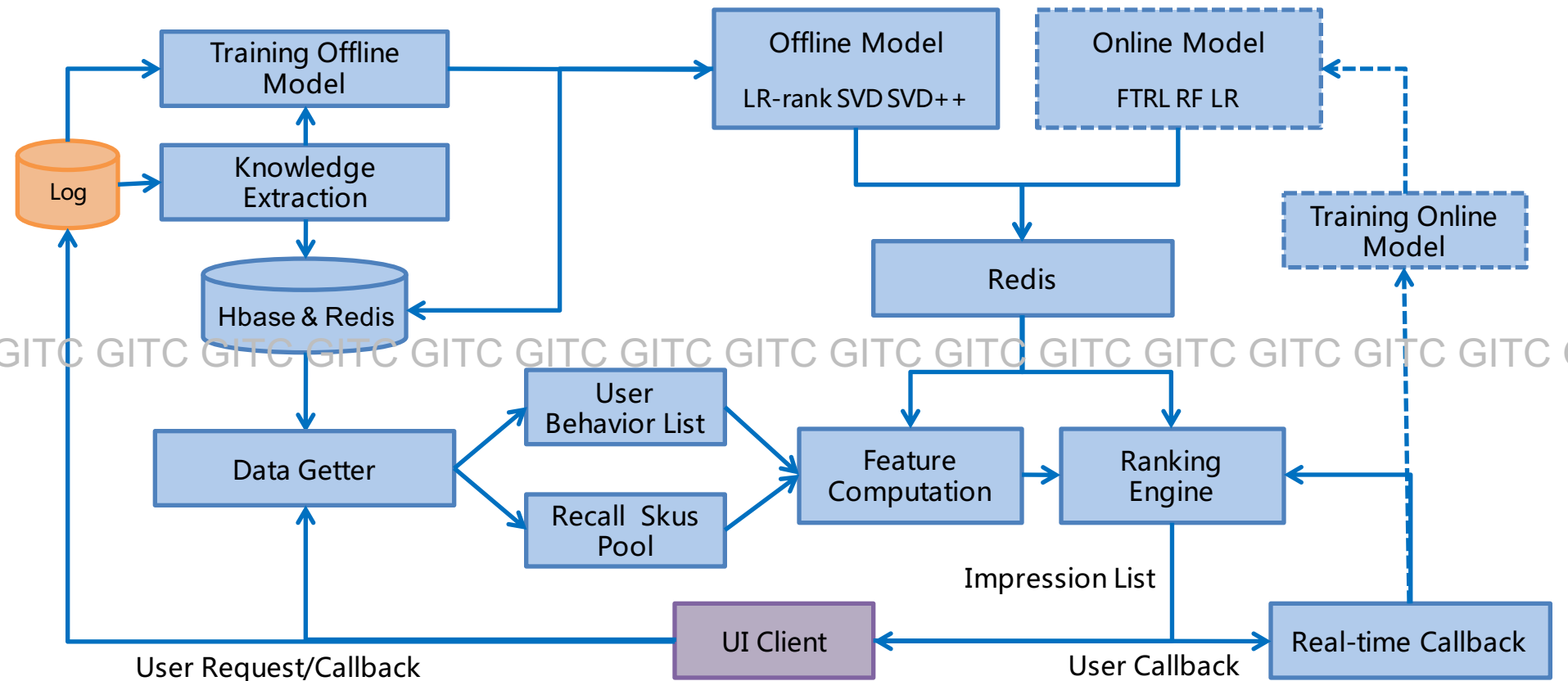
➤ 01

机器学习排序

➤ 02

推荐排序实践

推荐排序实践：架构



在线实时排序

- 过滤：业务需求（无库存、成人用品、促销商品）
 - 实时特征计算：人vs.商品，准实时行为特征
 - 多样性：多模型、多数据源融合
 - 新颖度
 - CTR/CVR
-
- 在线A/B测试

离线数据挖掘

- 隐语义模型数据：矩阵分解
 - 协同过滤：UserBasedCF、ItemBasedCF
 - 用户画像：品类、品牌、性别、年龄、购买力等
 - 商品画像：价格指数、适用人群和地理区域、购买周期等
 - 计算机视觉特征
-
- 离线测试



1. 收集用户行为，包括点击、加购、关注、下单等



2. 对行为进行过滤，比如：join(白名单)、统计截断、position-bias、多次加/删购等处理



3. 制定行为评分规则，生成评分矩阵



4. 训练矩阵分解模型



5. 导入缓存

- 用户因子向量 $p_u \in \mathbb{R}^f$ 和商品因子向量 $q_v \in \mathbb{R}^f$
- 评分偏置(bias) $b_{uv} = \mu + b_u + b_v$ (μ 是整个网站评分的均值, b_u 和 b_v 是用户/商品较网站评分的偏差)
- 用户对某商品 v 的预测评分为 $r'_{uv} = b_{uv} + p_u^T q_v$
- SVD++ 加入隐式反馈

$$r'_{uv} = b_{uv} + q_v^T \left(|R(u)|^{-\frac{1}{2}} \sum_{j \in R(u)} (r_{uj} - b_{uj}) x_j + |N(u)|^{-\frac{1}{2}} \sum_{j \in N(u)} y_j \right)$$

- 其中
 - $q_v, y_v, x_v \in \mathbb{R}^f$ 是商品的三个隐向量
 - $R(u)$ 代表被用户 u 打过分商品集合
 - $N(u)$ 代表用户隐式选择的商品集合 (即打分/未打分)

推荐排序实践：环境特征

某地高温
橙色预警

利用天气API定位某地高温橙色预警，该地区用户进入网站，首页猜你喜欢推荐空调、风扇等商品：

猜你喜欢—你的专属推荐

换一组



 海尔(Haier) KFR-23GW/12N WA13套机 小1匹P壁挂式定 ¥1799.00	 伊莱克斯(Electrolux)EAW25 FD13CA1 1匹P壁挂式定频 ¥1299.00	 美的(Midea) AC120-16B W空调扇 (冷风机 加湿 净化) ¥328.00	 美的(Midea) AC120-16BR W空调扇 (快速制冷 三档风) ¥429.00	 青岛啤酒 纯生 1000ml 24瓶 德国风味 企业自营销量保障 ¥110.00	 青岛啤酒 经典 500ml 12瓶 德国风味 企业自营销量保障 ¥69.00
---	---	---	--	--	--

某地雾霾
红色预警

利用天气API定位某地雾霾红色预警，该地区用户进入网站，首页猜你喜欢推荐空气净化器、口罩等商品：

猜你喜欢—你的专属推荐

换一组



 飞利浦(PHILIPS) AC4076 空气净化器 (智能空气测控) ¥2499.00	 格力大松(Gree) KJFC230 A-WG 空气净化器 ¥2579.00	 3M9001V防雾霾口罩9002V 带呼吸阀男女骑行PM2.5粉 ¥90.00	 美国 3M 9021 9022 KN90防颗粒物防尘口罩工业粉尘防 ¥110.00	 养生堂 天然维生素E软胶囊250mg*90粒 VE 维生素e ¥69.00	 养生堂天然维生素C咀嚼片850mg*80片 ¥82.00
---	--	--	---	---	--

推荐排序实践：深度学习与视觉特征



图像标注

对国美全站的商品图像进行标注：

1. 最主要的工作是数据清洗，即把不能和品类对应的图片删除或者重新进行品类校准
2. 统计国美全站最近一年内各品类下商品总数的分布；然后按照分布进行图片抽样



提取特征

图片的特征分为两个部分，一是通过深度学习得到的特征，二是图像色局部特征。

利用caffe训练CNN，将倒数第二层输出作为Feature Learning的结果提取出来
(Deep Learning + Transfer Learning)

利用局部特征算子(SIFT,kaze等)提取出图像的局部特征



降维

分别对前面两种特征进行降维处理：

用积量化 (Product Quantization) 的方法对CNN特征进行降维

用Fisher Vector对图像局部特征进行降维



匹配

采用最近邻搜索的方法找出每一个商品的相似商品集合

手工特征：

SIFT [Lowe 99]

Spin Images [Johnson&Herbert 99]

Textons[Malik et al. 99]

RIFT[Lazebnik 04]

GLOH[Mikolajczyk&Schmid 05]

HOG[Dalal&Triggs 05]

深度学习特征：



Zeiler&Furges, ECCV2014



推荐排序实践：提升推荐个性化及转化率

未经机器学习排序，网站首页猜你喜欢结果

猜你喜欢—你的专属推荐

换一组

					
乐扣乐扣 (300ML) 保温杯 ¥89.00	乐扣乐扣保温杯拉恩马克杯水杯 ¥122.00	膳魔师 (THERMOS) 保温保冷杯JNS- ¥172.00	THERMOS/膳魔师 高真空保温保冷 ¥175.00	乐扣乐扣保温壶保温杯水杯豪华 ¥239.00	富光 304 不锈钢保温杯 ¥38.00

使用机器学习提升推荐转化率

利用机器学习排序就是从数据中自动学习模式，在若干限定条件下，找出全局最优或者局部最优的近似值

利用机器学习排序后，网站首页猜你喜欢推荐商品

猜你喜欢—你的专属推荐

换一组

					
贝思嘉 生日蛋糕 摩尼巧克力 ¥99.00	贝思嘉 生日蛋糕 干果巧克力 ¥99.00	【团购价4100】西门子 ¥5088.00	凌思H440行车记录仪 高清夜视 ¥388.00	跑马人 冰激果三件套装 ¥79.00	妙厨宝 个性礼品冰激果 ¥109.00



Thanks!

