51CTO.com 技术成就梦想 | UBM

2012 / 10 / 25 ► 2012 / 10 / 26
中国 · 北京
October 25-26,2012
Beijing · China

World of Tech WOT

2012云计算架构师峰会

Cloud Computing Architects Summit China 2012

揭示企业级IT架构转型 分享最新技术的应用落地

主办单位：51CTO.com 技术成就梦想　联合主办：UBM

应用大数据能力

——当当网在个性化推荐&精准营销方面的探索

以“探索”为主线，让各位同学跟我一起亲身经历一次 2006年至今的当当网个性化推荐&精准营销 技术探索&架构革新之旅”

当当网
dangdang.com

时光穿越至2006年

设计模式 全部分类 搜索 高级搜索

计算机/网络

- 程序设计
  - C C++ C# VC VC++ (3982)
  - Java Java Script (2386)
  - 其他 (2272)
  - Basic VB VB Scri (1611)
  - .NET (1130)
  - ASP (789)
  - Pascal Delphi (493)
  - HTML XML (377)
  - 网站开发 (312)
- 图形图像 多媒体
- 计算机教材
- 家庭与办公室用书
- CAD CAM CAE

购买此商品的顾客也同时购买

大话设计模式
程杰 (作者)
★★★★½ (305)
平装
￥ 30.60

重构:改善既有代码的设计
福勒(Martin Fow...
★★★★½ (103)
平装
￥ 47.60

Head First设计模式(中文版)
弗里曼 (作者)
★★★★½ (96)
平装
￥ 66.80

代码大全(第2版)
史蒂夫•迈克康奈尔 (St...
★★★★½ (179)
平装
￥ 87.40

Effective C++:改善程序与设计的5 ...
梅耶(Scott Meye...
★★★★½ (70)
平装
￥ 45.50

Amazon.com Recommendations

Item-to-Item Collaborative Filtering

$$similarity(\vec{A}, \vec{B}) = \cos(\vec{A}, \vec{B}) = \frac{\vec{A} \bullet \vec{B}}{\|\vec{A}\| * \|\vec{B}\|}$$

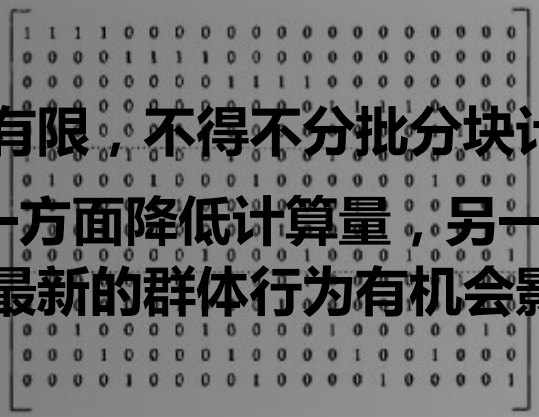
```
For each item in product catalog, I1
  For each customer C who purchased I1
    For each item I2 purchased by
       customer C
      Record that a customer purchased I1
       and I2
  For each item I2
    Compute the similarity between I1 and I2
```

当当网 dangdang.com

100万图书

1000万用户

双核
2G-4G内存
32位
50-200G磁盘



解决方案

- 多步骤处理：内存有限，不得不分批分块计算
- 只计算3-6个月：一方面降低计算量，另一方面数据较新，让用户最新的群体行为有机会影响推荐效果基本信息
- 剪枝：去除脏数据噪声、低频低销量冷门商品、高频畅销商品
- 稀疏矩阵的高压缩比的存储与支持高效查询仍然是问题

当当网
dangdang.com

P100	C001	C004	C008	C162	C589	C798

C001	C004	C008
P006	P001	P004
P100	P005	P009
P168	P100	P100
P457	P457	P235
P688		P688
		P889

稀疏矩阵的**高压缩比的存储**与**支持高效查询**解决方案

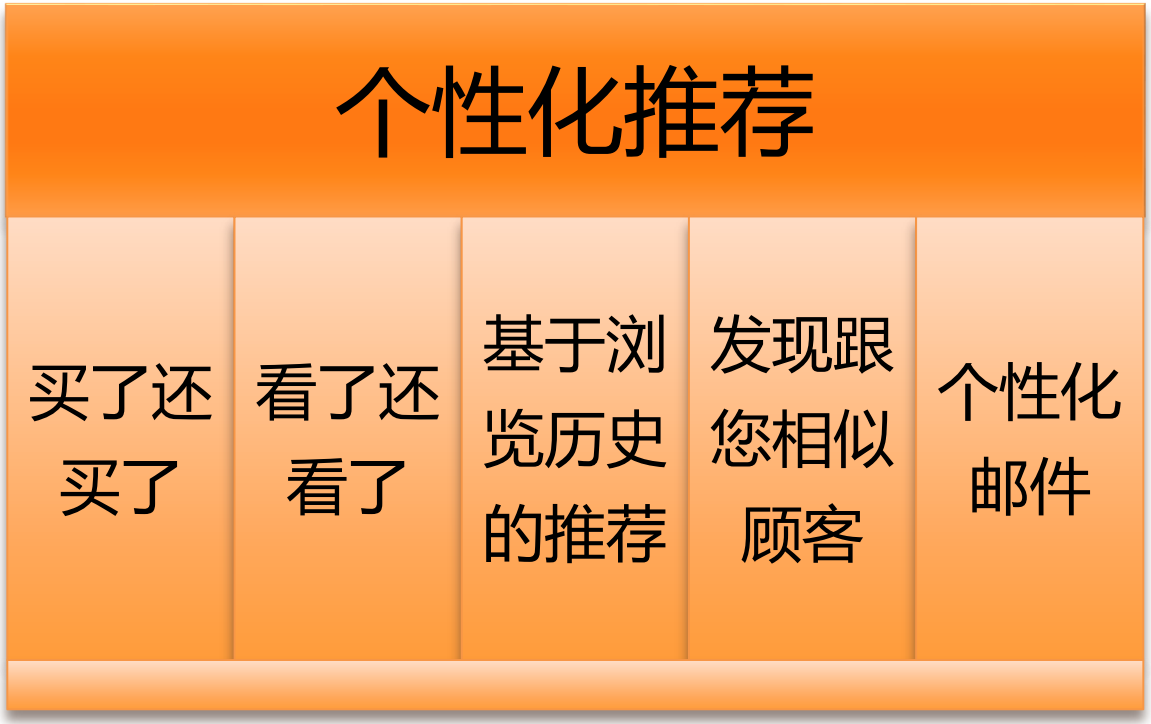
- 倒排索引
- 内存映射

空间：几十G -> 几百M

时间：处理全量数据2小时以内

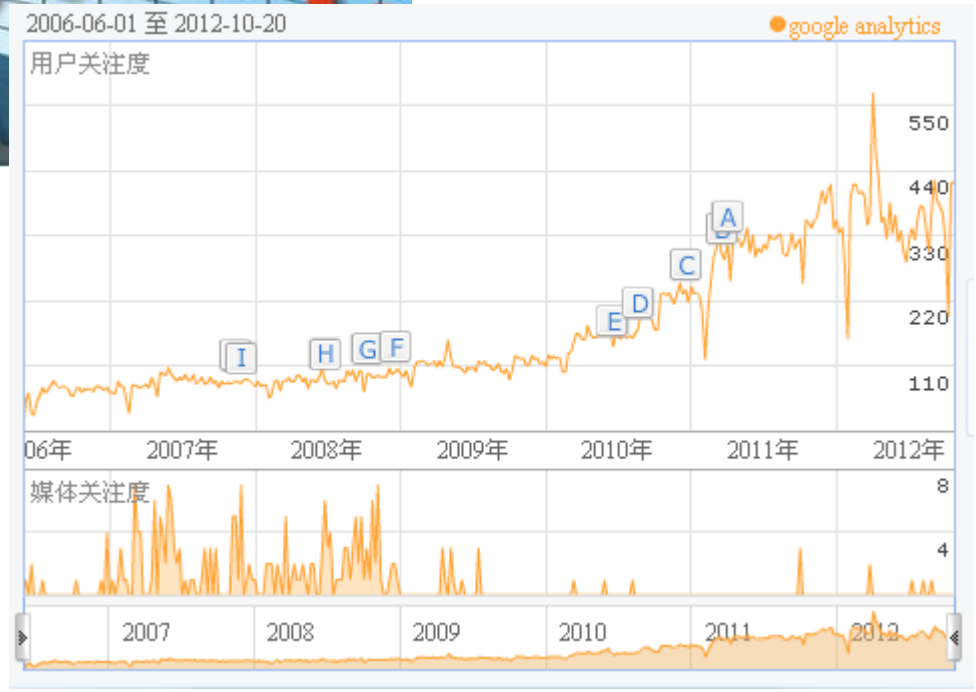
2006研发，2007上线。获得巨大成功！但可惜当时没数字证明☺

时光荏苒，2007、2008陆续推出基于c++的更多推荐产品☺

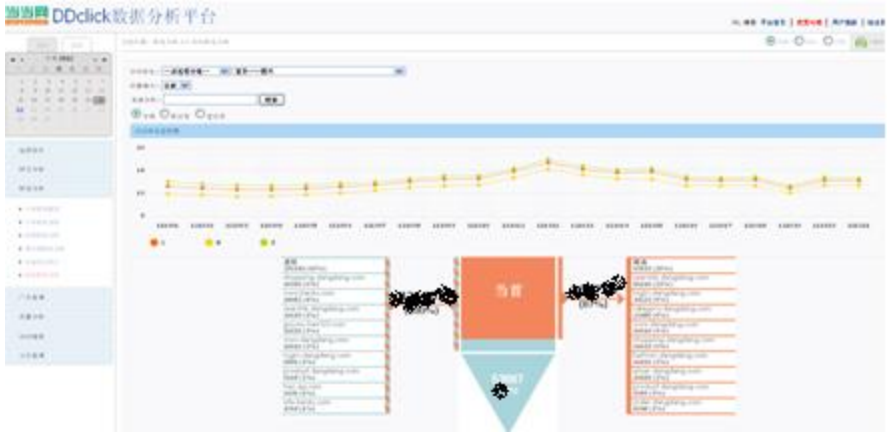


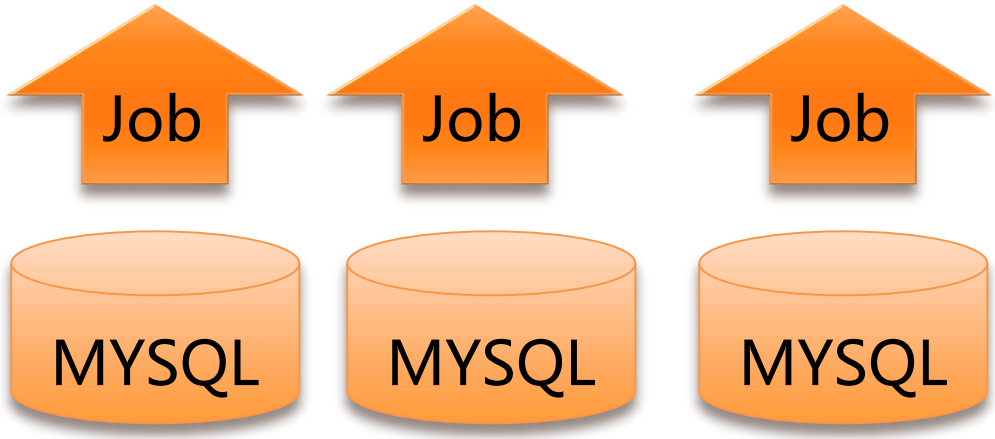
当当网 dangdang.com

时光穿越至2008、2009年



当当网
dangdang.com



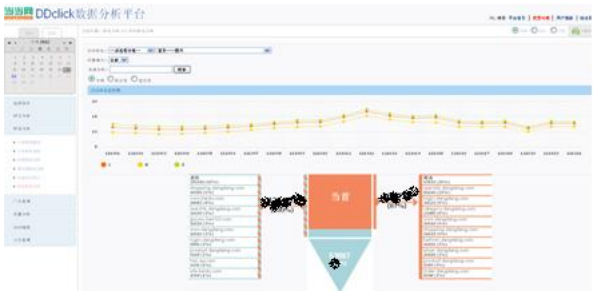


→ ?

随着时间的推移，2009、2010互联网各种新技术层出不穷：**hadoop、erlang、gearman**等等。这些新技术新思想不断对现有系统产生影响，并促成现有系统不断发展。精准营销生态系统进入新阶段。

神器！

$$similarity(\vec{A},\vec{B}) = \cos(\vec{A},\vec{B}) = \frac{\vec{A} \bullet \vec{B}}{\|\vec{A}\| * \|\vec{B}\|}$$



Jobs

A的频率

B的频率

AB的频率

hadoop

HIVE

Hadoop

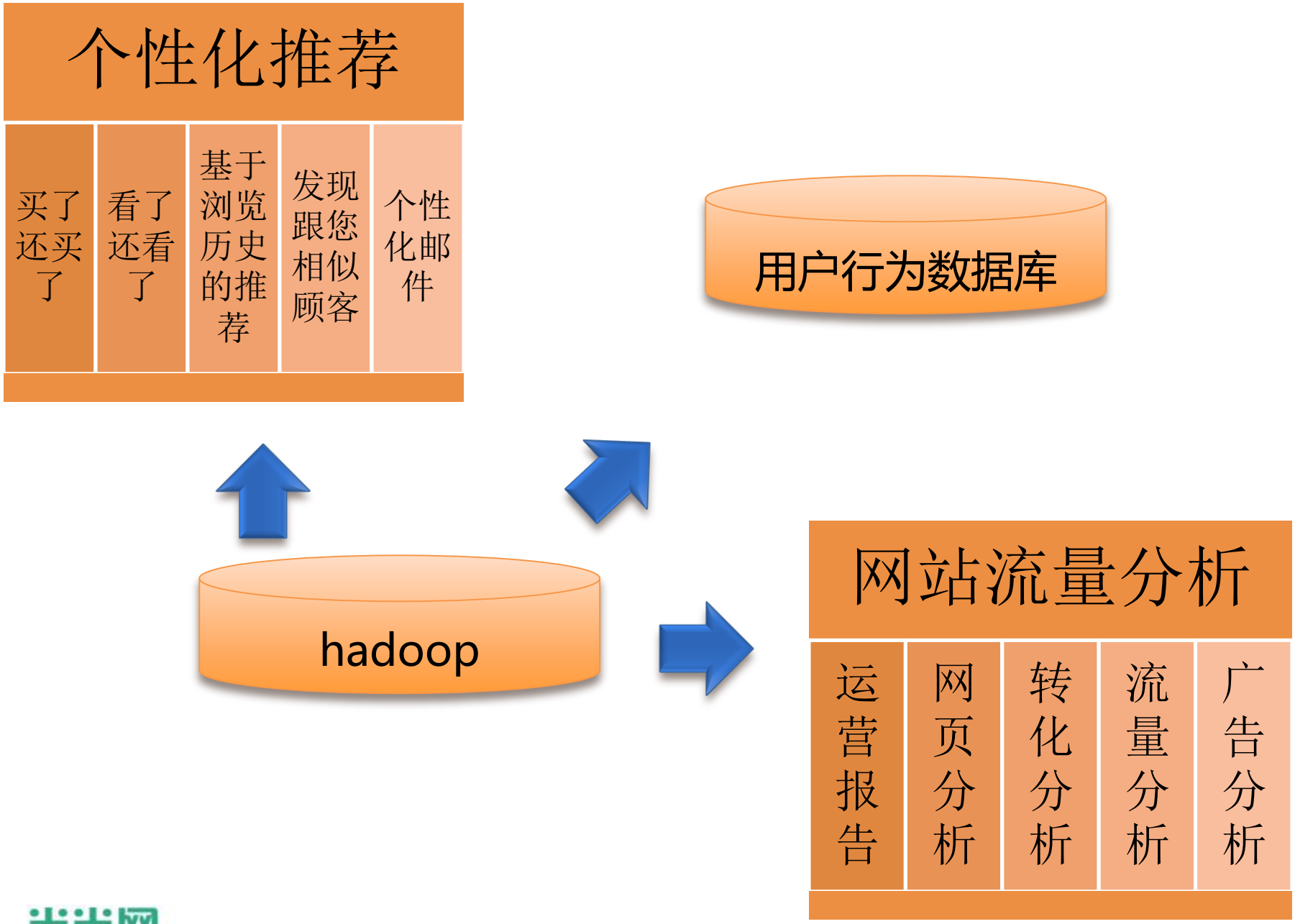
当当网 dangdang.com

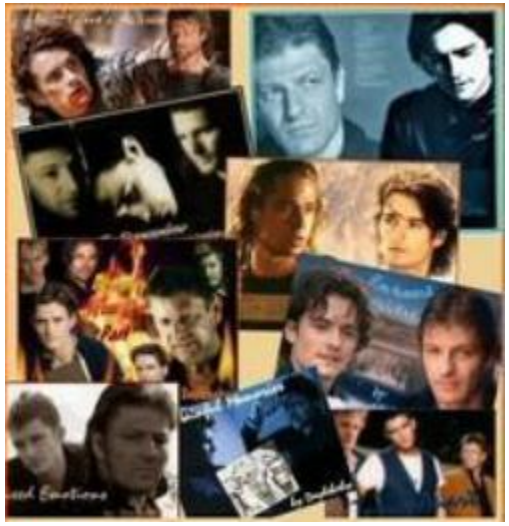
推荐/协同过滤

Non-distributed recommenders:
Taste(UserCF, ItemCF, SlopeOne)
Distributed Recommenders:
ItemCF

算法类	算法名	中文名
分类算法	Logistic Regression	逻辑回归
	Bayesian	贝叶斯
	SVM	支持向量机
	Perceptron	感知器算法
	Neural Network	神经网络
	Random Forests	随机森林
	Restricted Boltzmann Machines	有限波尔兹曼机
聚类算法	Canopy Clustering	Canopy聚类
	K-means Clustering	K均值算法
	Fuzzy K-means	模糊K均值
	Expectation Maximization	EM聚类（期望最大化聚类）
	Mean Shift Clustering	均值漂移聚类
	Hierarchical Clustering	层次聚类
	Dirichlet Process Clustering	狄里克雷过程聚类
	Latent Dirichlet Allocation	LDA聚类
	Spectral Clustering	谱聚类
关联规则挖掘	Parallel FP Growth Algorithm	并行FP Growth算法
回归	Locally Weighted Linear Regression	局部加权线性回归
降维/维约简	Singular Value Decomposition	奇异值分解
	Principal Components Analysis	主成分分析
	Independent Component Analysis	独立成分分析
	Gaussian Discriminative Analysis	高斯判别分析
进化算法	并行化了Watchmaker框架	
推荐/协同过滤	Non-distributed recommenders	Taste(UserCF, ItemCF, SlopeOne)
	Distributed Recommenders	ItemCF
向量相似度计算	RowSimilarityJob	计算列间相似度
	VectorDistanceJob	计算向量间距离
非Map-Reduce算法	Hidden Markov Models	隐马尔科夫模型
集合方法扩展	Collections	扩展了java的Collections类

Hadoop


个性化推荐
买了还买了
看了还看了
基于浏览历史的推荐
发现您似客顾相跟
个性化邮件
用户行为数据库
hadoop
网站流量分析
运营报告
网页分析
转化分析
流量分析
广告分析



姓名：程序猿

性别：男　　爱好：女

居住地：北京回龙观

行业：互联网

网购时间：22点-凌晨2点

身材：腹围 > 臀围 > 胸围

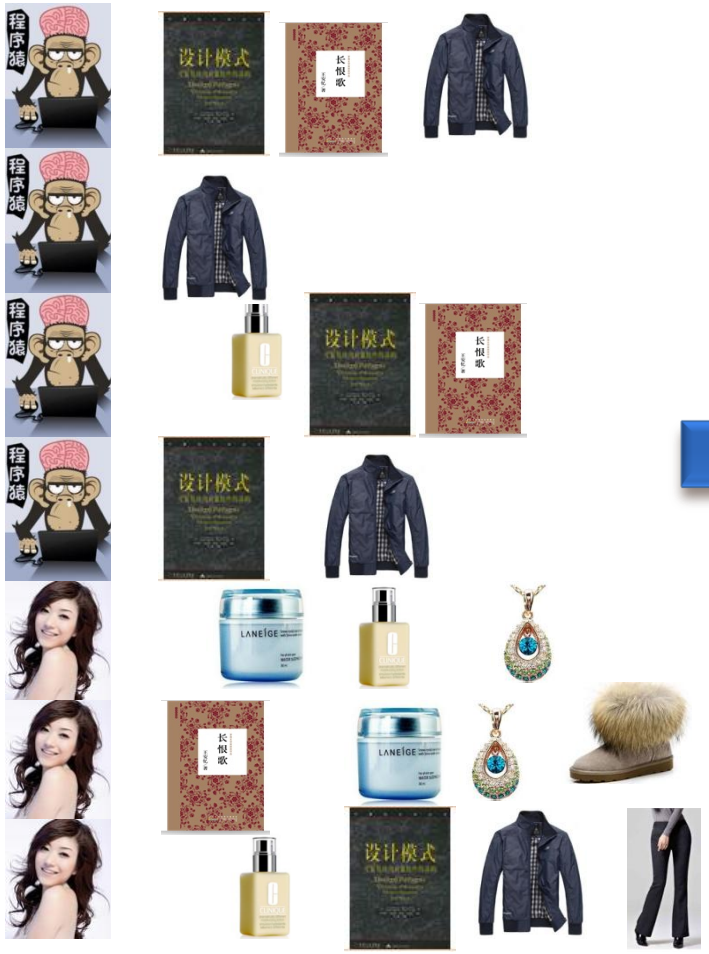
终端：chrome / Andriod

标签云：架构、高性能计算、分布式存储、重构、大数据处理、数据挖掘

用户行为数据库

当当网 dangdang.com

算法的力量



SVM

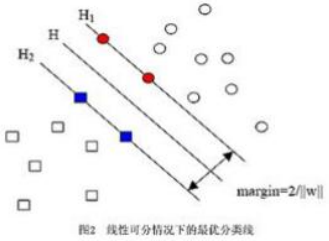
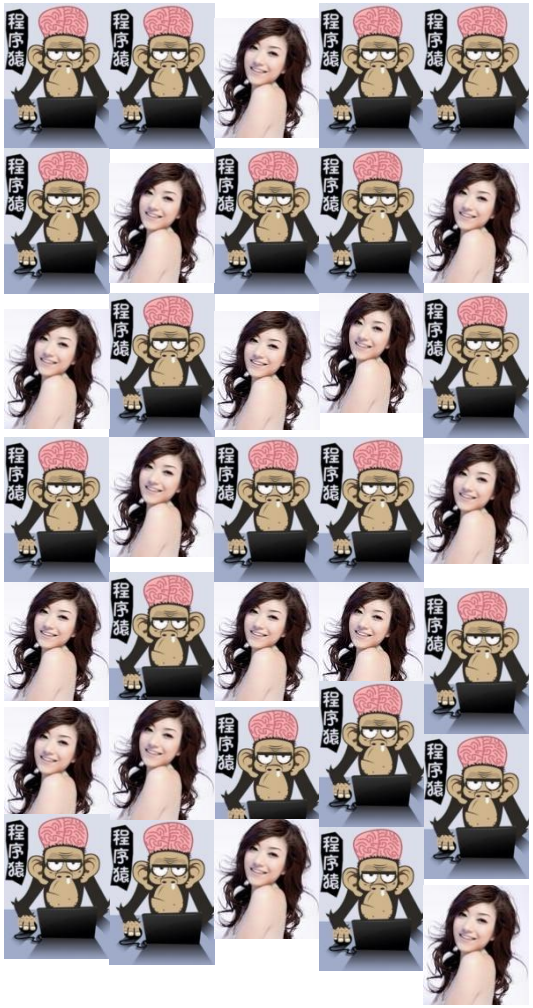
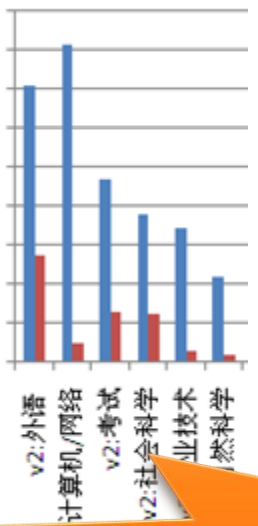


图2 线性可分情况下的最优分类线

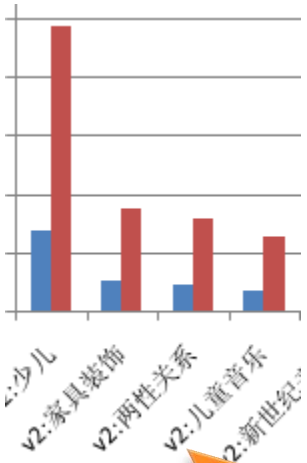


当当网
dangdang.com

最会被男性购买的图书分类



最会被女性购买的图书分类

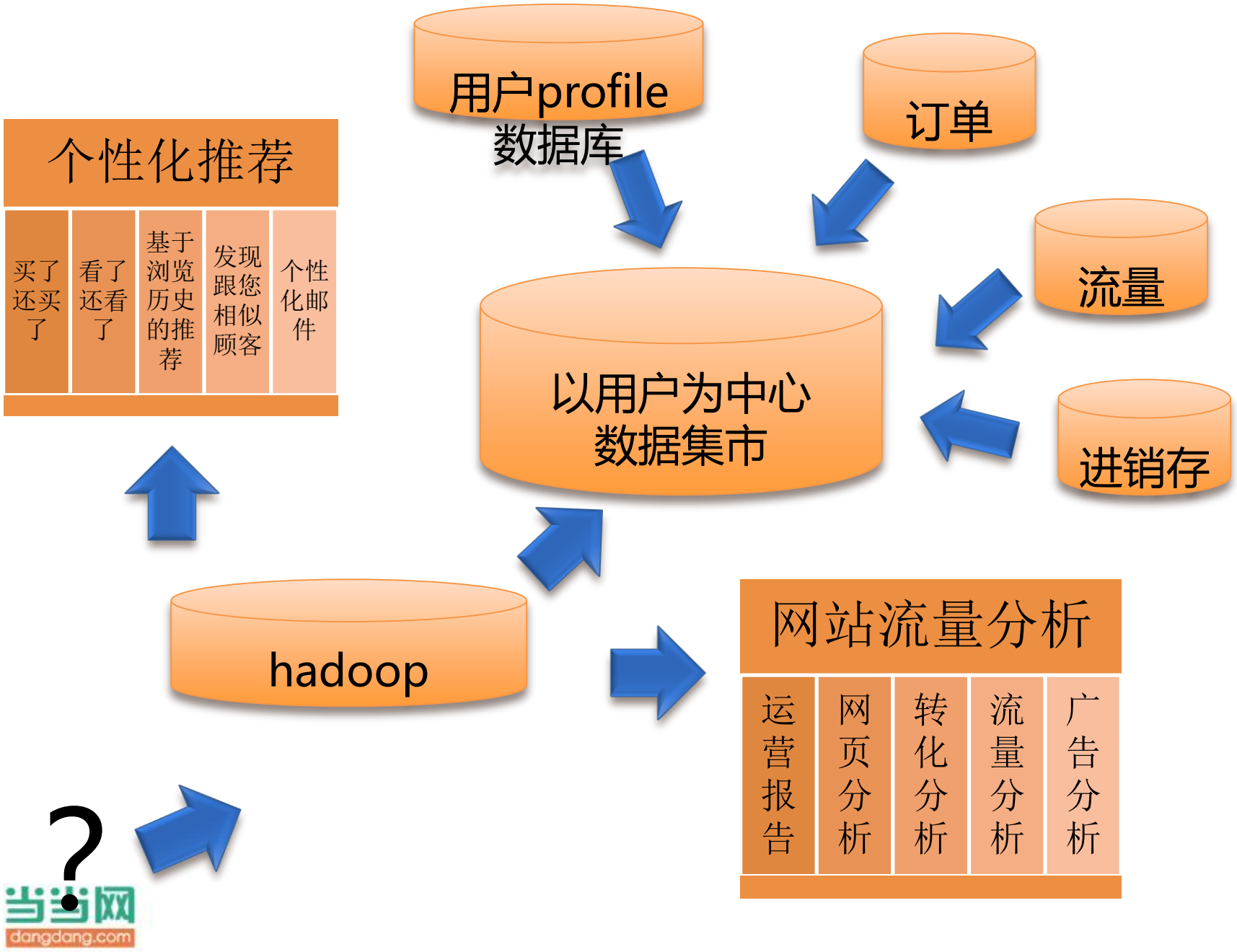


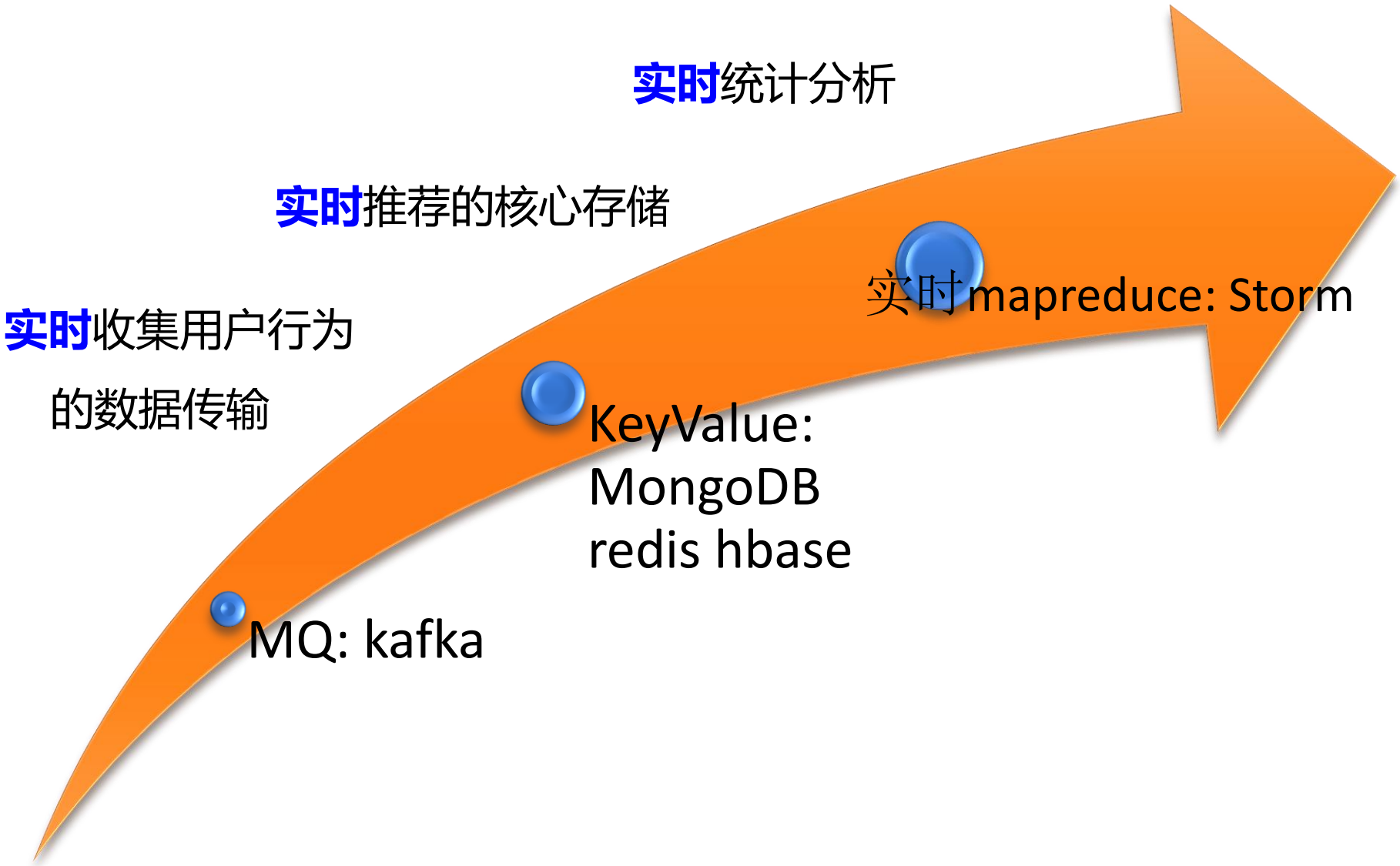
当当男是搞IT的很多

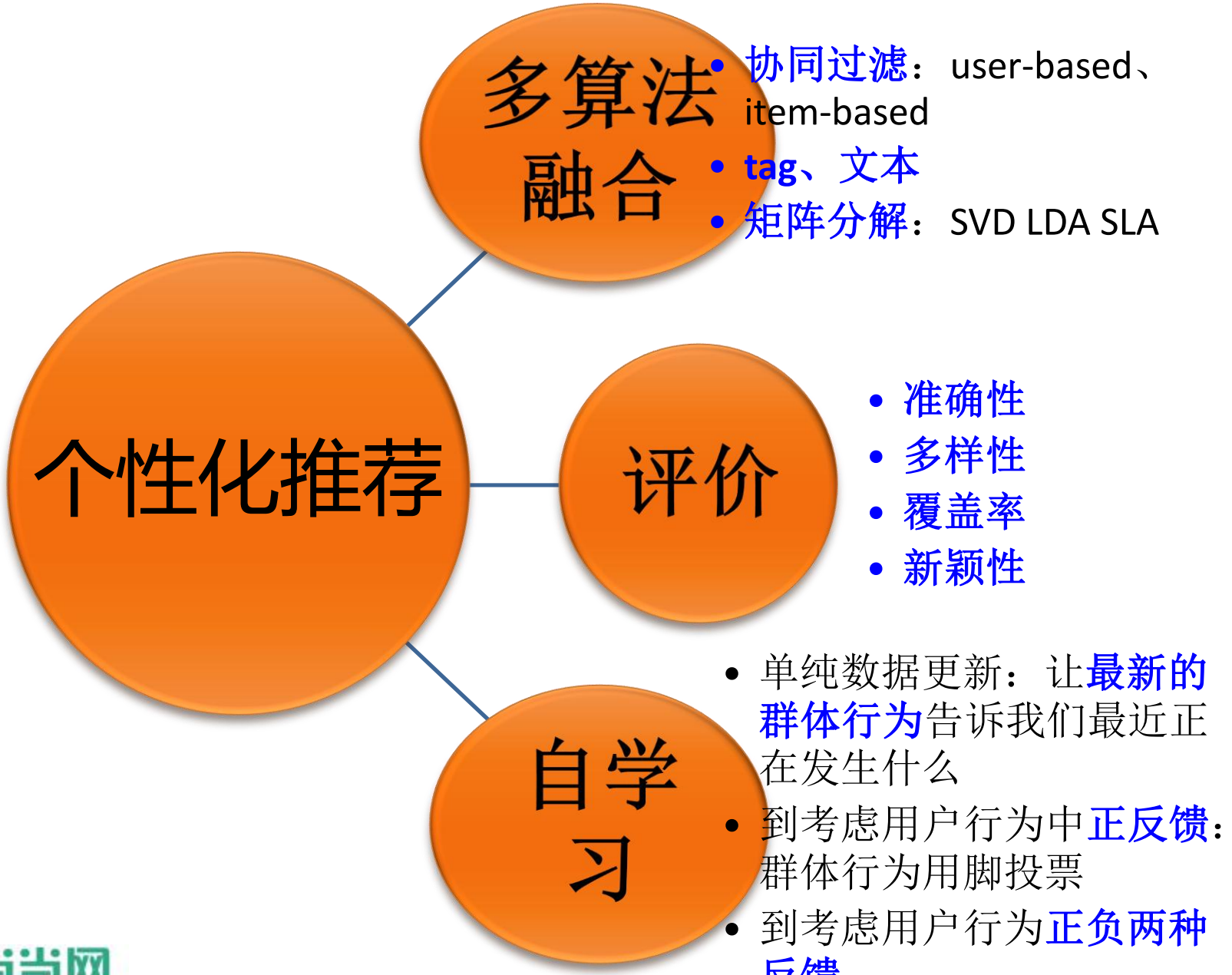
当当女喜欢看惊悚推理系列

当当女是望子成龙的好妈妈，经常买教辅

购买手链的当当男比当当女还多


用户profile
数据库
订单
流量
进销存
以用户为中心
数据集市
个性化推荐
买了还买了
看了还看了
基于浏览历史的推荐
发现您似客跟相顾
个性邮件
hadoop
网站流量分析
运营报告
网页分析
转化分析
流量分析
广告分析
?
当当网
dangdang.com





个性化推荐

- 买了还买了
- 看了还看了
- 基于浏览历史的推荐
- 发现您似跟相顾客
- 个性化邮件

网站流量分析

- 运营报告
- 网页分析
- 转化分析
- 流量分析
- 广告分析

重要模块

- ABtest
- 邮件平台
- 短信平台
- Anti-Fraud

大数据能力

- hadoop
- storm
- Mongodb Redis hbase
- kafka

算法能力

- 个性化推荐领域算法
- 自然语言处理算法
- 通用算法：聚类、分类、预测、回归等

用户数据集市

- 用户行为
- 订单
- 流量
- 进销存
- ERP

时时关注互联网最新技术动态、产品动态、业界动态，甚至国际大环境、国内外时事，这些因素或早或晚最终会影响到我们身处的行业和所负责的产品。如：站流量分析系统就是典型的例子、hadoop也是革命性的技术产品之一

时刻都在寻找下一块插图。打造自己的架构生态圈，其过程像拼图一样。每个组成模块，有自己的特点，专门解决合适的问题

Email:

fuqiang@dangdang.com

fqfuqiang@sina.com

新浪微博：

@fq傅强

当当网
dangdang.com