# 推荐系统 – 介绍和交流

1.0 -> 2.0

吴源林
✉ wyl2000@gmail.com ✆186 0219 5030
http://v2000.info

# 目录

- 推荐系统1.0
- 推荐系统2.0
  - 推荐算法原理
  - 推荐系统：架构，软、硬件
  - 如何向用户进行推荐 – 营销、运营和交互设计
  - 推荐效果评估

# 推荐系统1.0

# 购物篮分析

- A、B两家餐厅被客户预定的情况（人数）

| | A餐厅 | 非A餐厅 | 总计 |
|---|---|---|---|
| B餐厅 | 15 | 5 | 20 |
| 非B餐厅 | 75 | 5 | 80 |
| 总计 | 90 | 10 | 100 |

支持度（｛A餐厅｝）= 90
支持度（｛B餐厅｝）= 20
支持度（｛A餐厅，B餐厅｝）= 15
可能性（｛A餐厅｝）= 90/100 = 0.9
可能性（｛B餐厅｝）= 20/100 = 0.2
可能性（｛A餐厅，B餐厅｝）= 15/100 = 0.15

可能性（A餐厅|B餐厅）=15/20=0.75　吃过B餐厅吃A餐厅的可能性
可能性（B餐厅|A餐厅）=15/90=0.167　吃过A餐厅吃B餐厅的可能性

# 缺点

- 餐厅-客户个性不能被体现
- 预订少的餐厅容易被剔除在外
- 马太效应：热门餐厅获得推荐机会更多
- 针对单个会员，推荐时机及餐厅变化相对少

# 推荐系统2.0

# 协同过滤算法

**协同过滤**（Collaborative Filtering），利用某兴趣相投、拥有共同经验之群体的喜好来推荐使用者感兴趣的资讯。

- 收集使用者资讯
- 针对用户/项目的最近邻搜索

  例如：要对餐厅A 和餐厅 B 进行相似性计算，要先找出同时对 A 和 B 打过分的组合，对这些组合进行相似度计算

- 产生推荐结果

  利用会员的偏好，接合群体智慧，进行推荐

# Item Based 算法

| | 101 | 102 | 103 | 104 | 105 | 106 | 107 | | U3 | | R |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 101 | 5 | 3 | 4 | 4 | 2 | 2 | 1 | | 2.0 | | 40.0 |
| 102 | 3 | 3 | 3 | 2 | 1 | 1 | 0 | | 0.0 | | 18.5 |
| **103** | **4** | **3** | **4** | **3** | **1** | **2** | **0** | **X** | **0.0** | **=** | **24.5** |
| 104 | 4 | 2 | 3 | 4 | 2 | 2 | 1 | | 4.0 | | 40.0 |
| 105 | 2 | 1 | 1 | 2 | 2 | 1 | 1 | | 4.5 | | 26.0 |
| 106 | 2 | 1 | 2 | 2 | 1 | 2 | 0 | | 0.0 | | 16.5 |
| 107 | 1 | 0 | 0 | 1 | 1 | 0 | 1 | | 5.0 | | 15.5 |

餐厅亲密度矩阵：两两餐厅，某种行为发生的频次（频次越大，则暗含这两家餐厅存在某种相似性越高）

会员偏好向量：对所有餐厅的偏好度（如点评，下单数，访问数）

预测推荐：用亲密度矩阵和偏好向量进行计算

# 协同过滤算法

**优点**

- 解决机器难以自动进行内容分析的资讯
- 共用其他人的经验
- 推荐新资讯
- 个性化，自动化程度较高

**缺点**

- 新使用者问题(New User Problem) 系统开始时推荐品质较差
- 新项目问题(New Item Problem) 品质取决于历史资料集
- 稀疏性问题（Sparsity）
- 系统延伸性问题（Scalability）

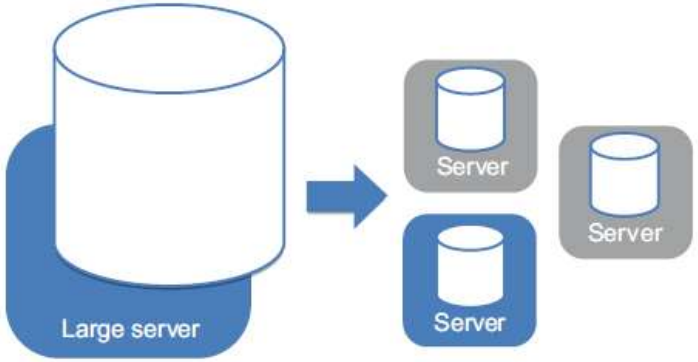
# 如何推荐？

- 推荐体验设计

除推荐物品的展示外，还有 **推荐理由和反馈！**


- 可能的推荐时机：
  - 网站|移动客户端|呼叫中心|EDM

# 运算系统：单点 -> 分布




hadoop

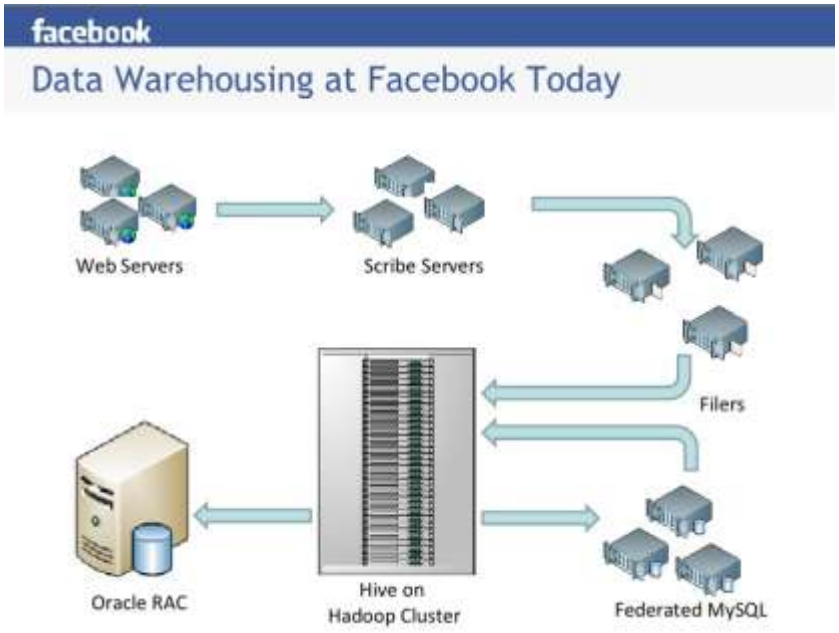
HIVE

Java

mahout

CentOS

# 行业应用案例：Facebook





FB日均处理：

- 25亿 Facebook上分享的内容条数
- 27亿 “赞”的数量，
- 3亿 上传照片数
- 500+TB 新产生的数据
- 105TB 每半小时通过Hive扫描的数据
- 100+PB（1PB=1024TB）单个HDFS（分布式文件系统）集群中的磁盘容量

“大数据的意义在于真正对你的生意有内在的洞见。如果你不能好好利用自己收集到的数据，那你只是空有一堆数据而已，不叫大数据。”

http://news.cnet.com/8301-1023_3-57498531-93/facebook-processes-more-than-500-tb-of-data-daily/

# DEMO 演示

# 数据源和同步

- 订单
- 点评
- 点击流
- GPS信息
- ……


数据同步

- Microsoft Sync Framework
- FTP

# 效果评估

- 可能的商业产出指标：
  - 财务：点击率，转化率，订单增长
  - 访问行为：点击量，访问深度，与搜索或分类目录比较
  - 体验：用户满意度，[净推荐者值（Net Promoter Score）]


| | 网站搜索状态 | 订餐转化指标 (目标 1 的转化率) | 访问次数 |
|---|---|---|---|
| 1. | Visits With Site Search | | |
| | 相关餐厅 | 19.82% | 0.39% |
| | 所有访问次数 | 20.23% | 11.84% |
| 2. | Visits Without Site Search | | |
| | 相关餐厅 | 8.41% | 0.85% |
| | 所有访问次数 | 2.39% | 88.16% |

# 效果评估

- 方法：实验、调查和AB测试
  - 预测准确度
  - 覆盖率（对推荐餐厅长尾的发掘能力）
  - 多样性
  - 新颖性
  - 惊喜度
  - 信任度
  - 实时性
  - 健壮性

- 缺陷：
  - 效果评估可能很难做到逻辑完善、结果公正，欠缺公认的定论

# 推荐阅读

迎接 个 性 化 时 代 浪潮


购 买　样章试读

## 推荐系统实践

浪潮之巅作者吴军作序推荐

序言　作者简介　目录

推荐在今天互联网的产品和应用中被广泛采用，包括今天大家经常使用的相关搜索、话题推荐、电子商务的各种产品推荐、社交网络上的交友推荐等。但是，至今还没有一本书从理论上对它进行系统地分析和论述。《推荐系统实践》这本书恰恰弥补了这个空白。

该书总结了当今互联网主要领域、主要公司、各种和推荐有关的产品和服务，包括：

亚马逊的个性化产品推荐；

Netflix的视频和DVD推荐；

Pandora的音乐推荐；

Facebook的好友推荐；

Google Reader的个性化阅读；

# 谢　谢

吴源林
✉ wyl2000@gmail.com ✆186 0219 5030
http://v2000.info