面向广告主的推荐

百度　江申

Oct. 27 2012

Outlines

I. 背景知识介绍
   1. 互联网广告
   2. 百度面向广告主的推荐产品
II. 设计要点
   1. 分析清楚技术目标
   2. 充分挖掘来自数据与人的信息
   3. 合理利用反馈信息
   4. 设计合适的推荐理由
   5. 注意评估结果的显著性
III. 一些经验
IV. Q&A

Baidu 百度

背景知识—搜索广告

Bai du 百度

背景知识—搜索广告

投放系统如何进行广告投放？

Query: 查询关键词
Bidword：广告主设置的推广关键词

网民输入Query → 投放系统找到与Query（及网民）相关且满足投放设置的广告（一条广告包含**拍卖词Bidword，出价，创意**等信息） → 投放系统通常根据每条广告的“预估点击率*出价”排序，投放前N个广告。

“拓荒保洁公司” →
（“保洁公司”，0.5，创意1)
（“开荒保洁公司”，0.8，创意2)
（“拓荒保洁”，1.5，创意3)
（“新家保洁”，1.6，创意4)
……
→
（“拓荒保洁”，1.5，创意3)→0.7*1.5=1.05
（“新家保洁”，1.6，创意4) →0.6*1.6=0.96
（“保洁公司”，0.5，创意1) →0.7*0.5=0.35

背景知识—搜索广告

广告主（或委托账户管理员）如何进行投放设置？

(1) 建立推广账户(account)

一个账户(account)包含多个推广计划(campaign)

一个推广计划(campaign)包含多个推广单元(group)

(2) 通过一系列设置来表达投放需求与策略

a. 添加拍卖词(bidword)

b. 写广告创意

c. 设定拍卖词出价

d. 设定拍卖词匹配方式

e. 设定预算

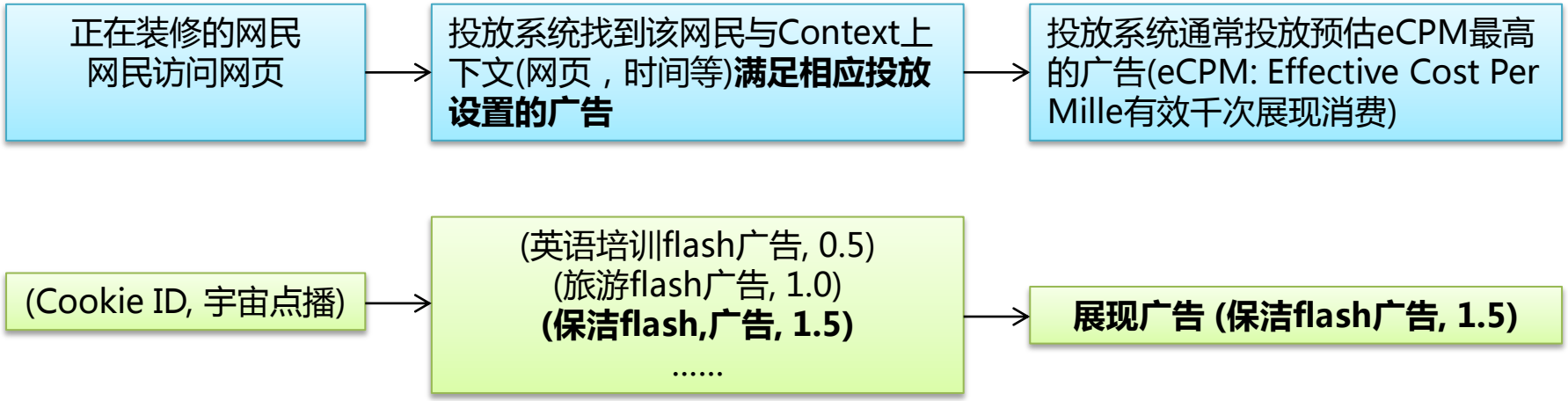
……

背景知识—展示广告


首页
电影
电视剧
综艺娱乐
经典动漫
纪录片
公开课
排行榜
全 部
百度影音点播
快播使用帮助
全部分类：
动作片
喜剧片
爱情片
科幻片
恐怖片
剧情片
战争片
国产剧
港台剧
日韩剧
欧美剧
马泰剧
高清频道
粤语经典
北京宏远易通清洗服务有限公司
24小时服务电话:
010-68474647 010-68471522
管道疏通
管道清洗
蹲坑去碱
管道安装维修
清掏化粪池
马桶疏通维修
★提供优质保洁服务★
成品保护、保洁托管
石材养护、石材结晶
北京世纪飞翔物业管理有限公司
TEL:010-85846188
您现在的位置： 首页 >> 欧美剧 >> 危机边缘第1季
《危机边缘第1季》
主演： Kirk Acevedo Blair Brown Joshua Jackson
导演：
类型： 欧美剧
地区： 欧美
语言： 英语
年份：
评分： 0.0分(0人评分)
播放： 120
我要收藏 我要分享 我要报错
我要评分：
0 顶一下 0 踩一下
下 载 快车下载 BT下载
FRINGE
卫生环境，幸福生活
承接 烟道清洗 石材护理
电话：010-67625036

背景知识—展示广告

B. 投放系统根据投放设置进行广告投放：



背景知识—展示广告

A. 广告主（或委托的账户管理员）进行投放设置：

(1) 建立推广账户(account)

(2) 通过一系列设置来表达投放需求与策略

a. 设置创意（图片或文字）
b. 设置出价（CPC/CPM/CPS）
c. 设置投放网站
b. 设置主题关键词
c. 设置投放地域
d. 设置投放人群(性别，年龄等)
……

Baidu 百度

背景知识—为什么要做面向广告主的推荐

互联网广告是什么？

一场广告主，媒体，网民之间共赢的利益交换游戏。

广告主现状如何？

中国广告主的成熟程度总体较低，SEM行业不如美国繁荣。

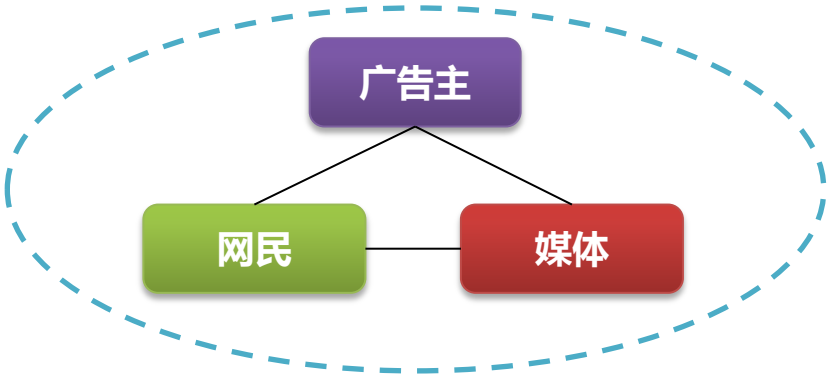
为什么需要面向广告主的推荐？

帮助广告主更好地玩好游戏。

广告系统来做比SEM公司来做的优势？

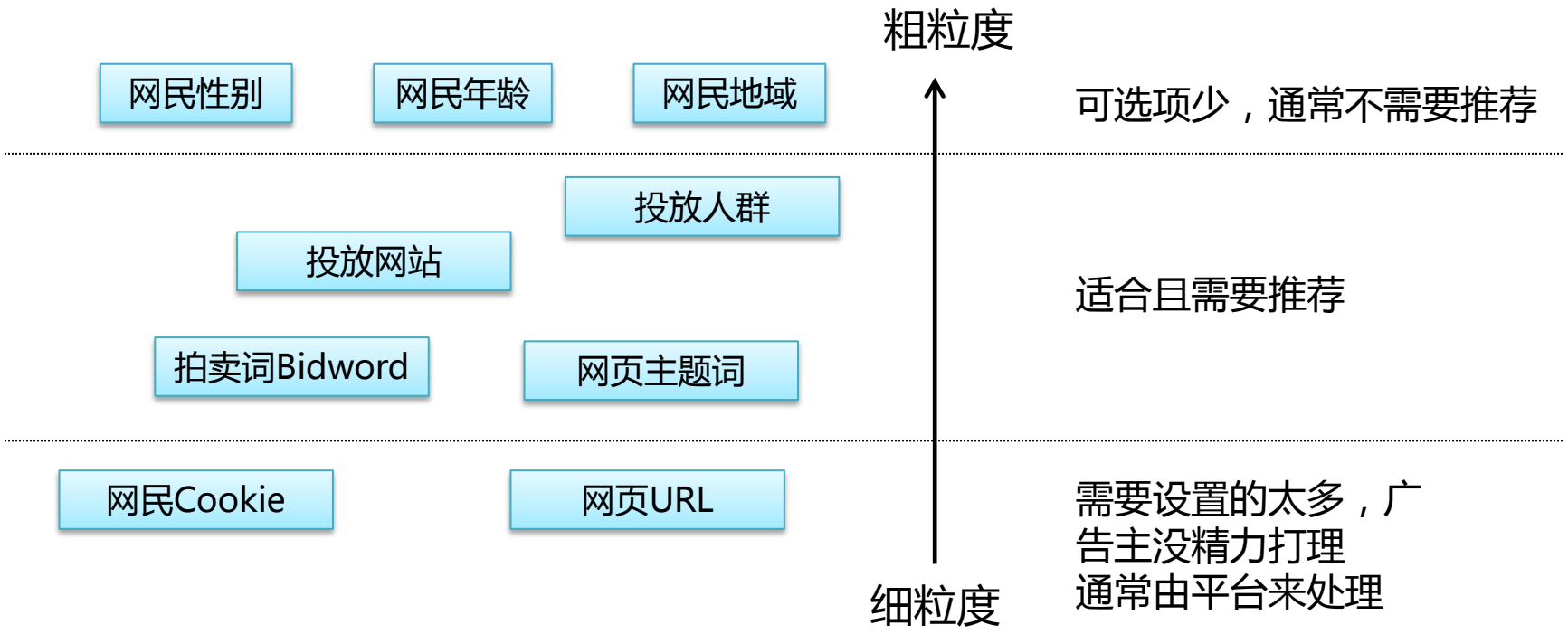
拥有账户数据或/与投放历史数据

互联网广告生态圈



背景知识—百度面向广告主的推荐产品

广告主表达投放需求与投放策略的基本元素：



背景知识—百度面向广告主的推荐产品

- 凤巢（搜索广告）
  - 拍卖词推荐
    - **主动推荐**（不带query推荐）
    - **被动推荐**（带query推荐）
    - 等等
  - 出价(Bidding)建议
  - 匹配模式建议
- 网盟（展示广告）
  - 主题词推荐
  - 网站推荐

拍卖词推荐在百度的作用

Fact 1:

推荐后被采用的拍卖词带来的收入占总体收入的约 **50%** 左右。

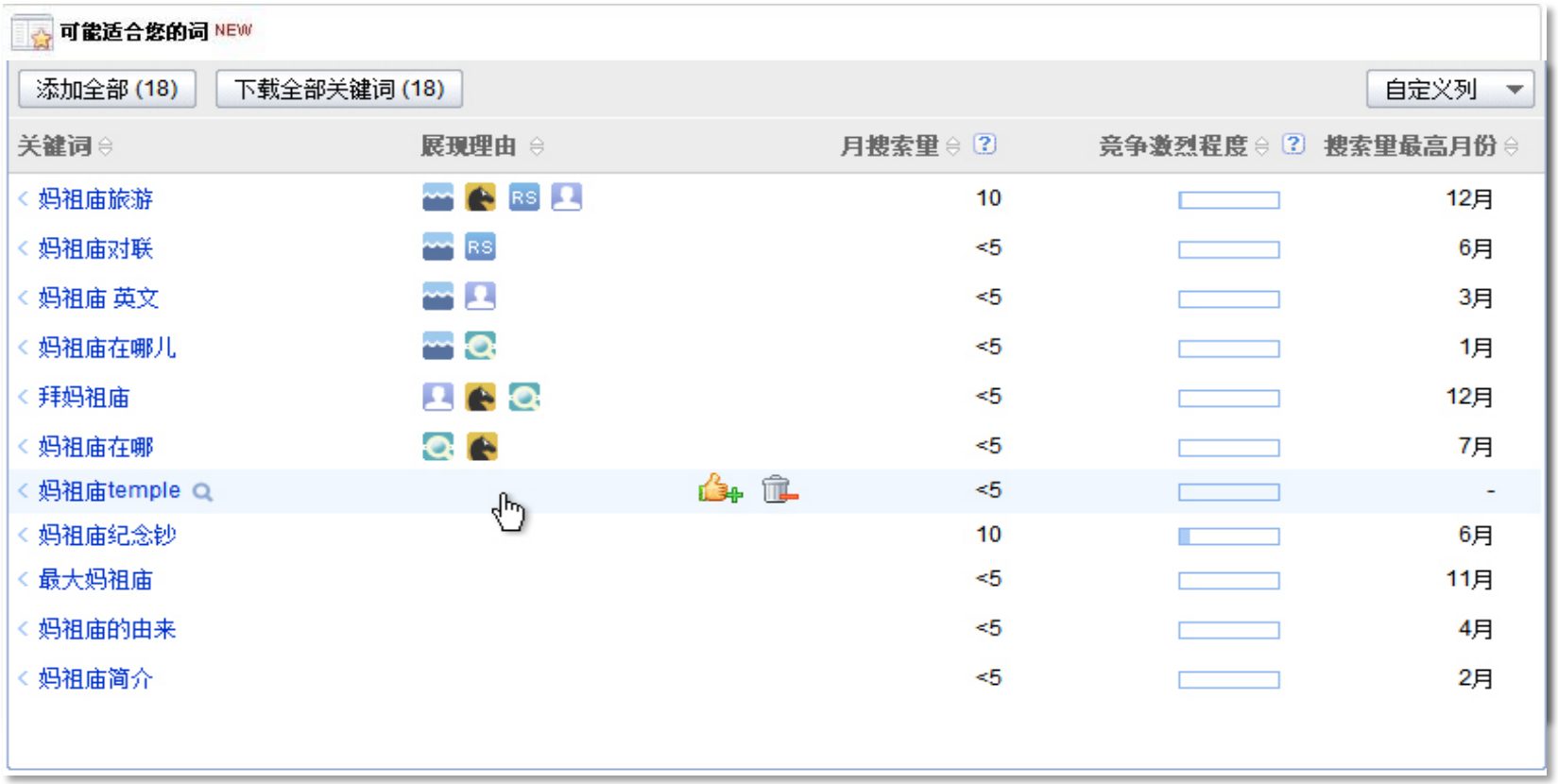
冷静，这是一个upper bound。

Fact 2:

推荐后被采用的拍卖词中能为广告主带来流量的词占例，是平均水平的约 **2** 倍。

背景知识—凤巢拍卖词主动推荐

拍卖词主动推荐：input是 (广告主id, context)



Baidu 百度

背景知识—凤巢拍卖词被动推荐

拍卖词被动推荐：input是 (Query, 广告主id，context)

Where are we?

I. 背景知识介绍
   1. 互联网广告
   2. 百度面向广告主的推荐产品

II. **设计要点**
   1. 分析清楚技术目标
   2. 充分挖掘来自数据与人的信息
   3. 合理利用反馈信息
   4. 设计合适的推荐理由
   5. 注意评估结果的显著性

III. **一些经验**

IV. **Q&A**

Baidu 百度

设计要点—分析清楚技术目标

需要考虑“推荐被采用以后的效用”：

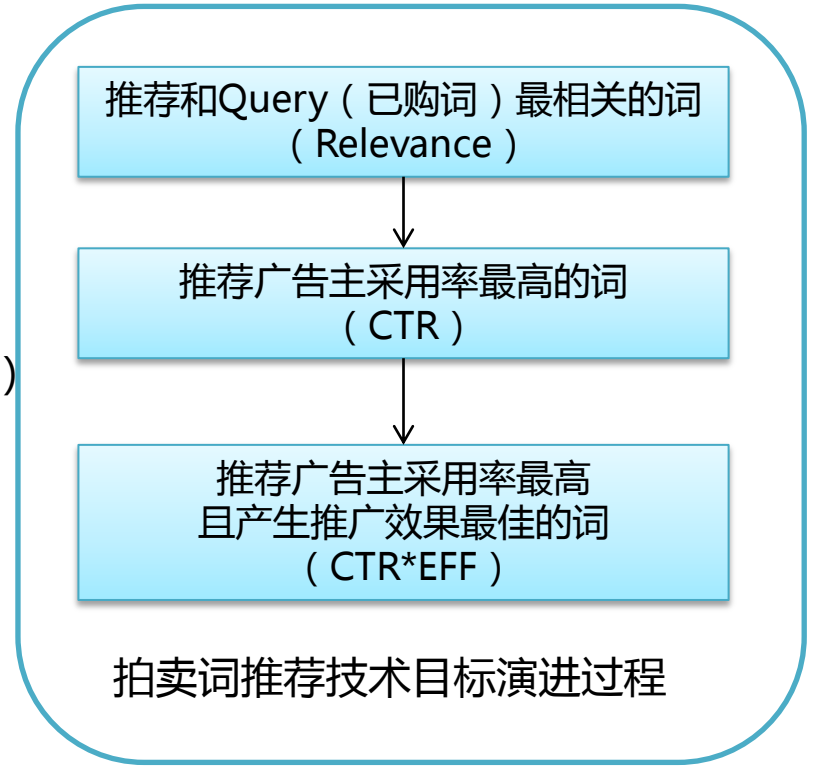
（1）是与大多数推荐系统的主要区别

（2）是广告投放系统做推荐的最大优势

用户产品与商业产品中的推荐侧重点：

前者：用户体验指标（Novelty, Serendipity等）

后者：商业利益（商业指标如CPC, ROI等）



拍卖词推荐技术目标演进过程

Where are we?

I. 背景知识介绍
   1. 互联网广告
   2. 百度面向广告主的推荐产品
II. 设计要点
   1. 分析清楚技术目标
   2. 充分挖掘来自数据与人的信息
   3. 合理利用反馈信息
   4. 设计合适的推荐理由
   5. 注意评估结果的显著性
III. 一些经验
IV. Q&A

Baidu 百度

算法工程师在做什么

在学校的时候以为：

标准算法1 → 算法1改进版 → 标准算法2 → 算法2改进版 → …

进公司以后发现：

数据A：定制算法1 → 数据A+数据B：定制算法1 → 调参数

→ 数据A+数据B：定制算法1+定制算法2 → 加规则 → …

Baidu百度

设计要点—充分挖掘来自数据与人的信息

策略架构思路：

建设可以很好融合各种数据的系统（数据知识），

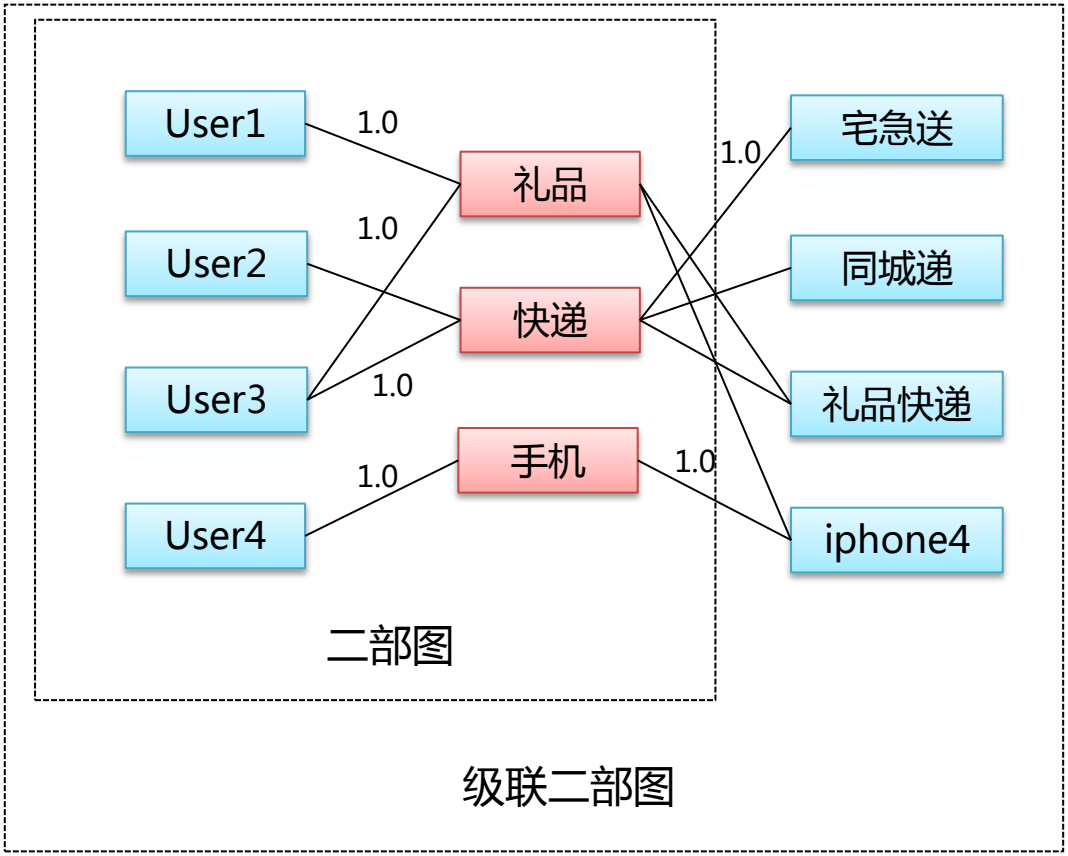
建设容易根据对数据的理解调整算法的系统（人的知识）

推荐引擎设计：

基于级联二部图的分布式挖掘框架（Based on Hadoop）

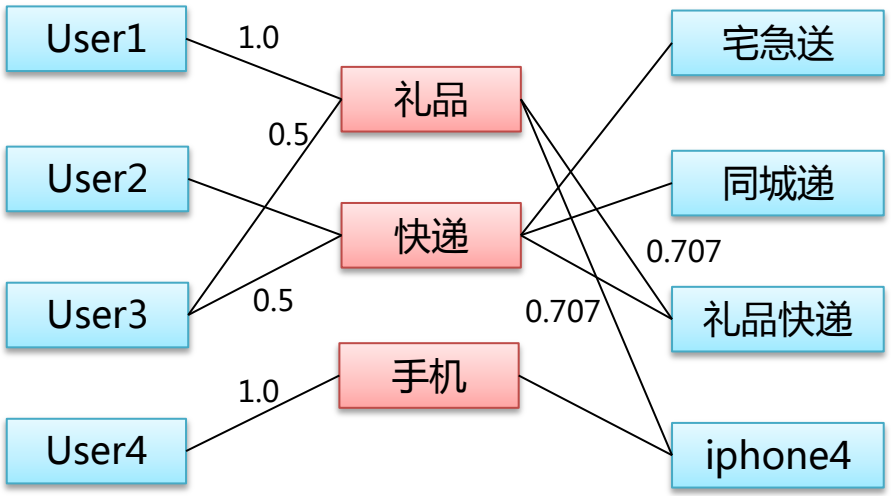
Bai du 百度

什么是级联二部图？



Note: 二部图有时也可以当成左右结点相同的级联二部图来看

挖掘步骤一



1. 左右结点权重调整 (通常是归一化)

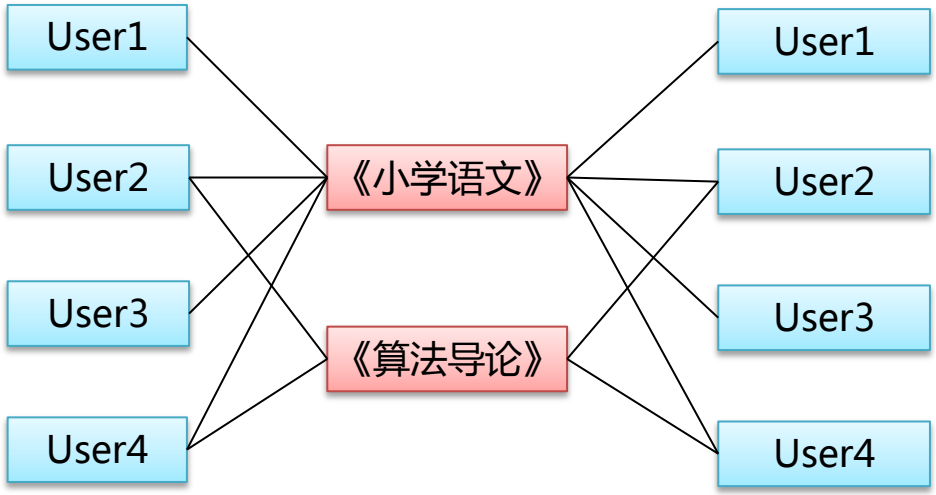
对于每个左结点（右节点），调用**模块Norm_A(Norm_B)**归一化所有与之相连的边的权重。

Norm_A 输入：某个左结点所有边的权重

Norm_A 输出：该左结点所有边的**归一化后的权重**

Norm_A例子：（1）L1-Norm: 权重和为1；（2）L2-Norm: 权重平方和为1

挖掘步骤二



2. 中间结点通过惩罚

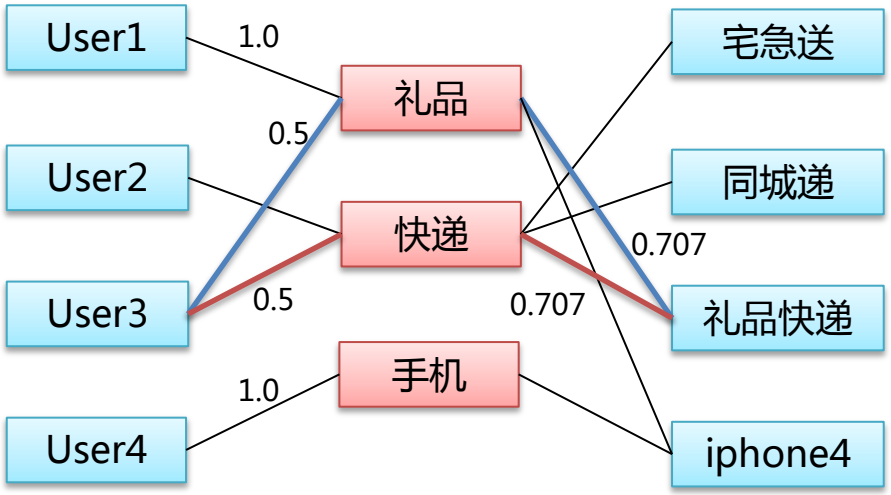
类似tf-idf中idf的作用。通过调用**模块Punish_Method**, 对每个结点计算一个惩罚值。

Punish_Method输入：中间结点的左侧所有边权重，右侧所有边权重

Punish_Method输出：中间结点的**通过惩罚值**

Punish_Method例子：通过惩罚值 = 1/（左侧所有边权重和）

挖掘步骤三



3. 单条路径权重计算

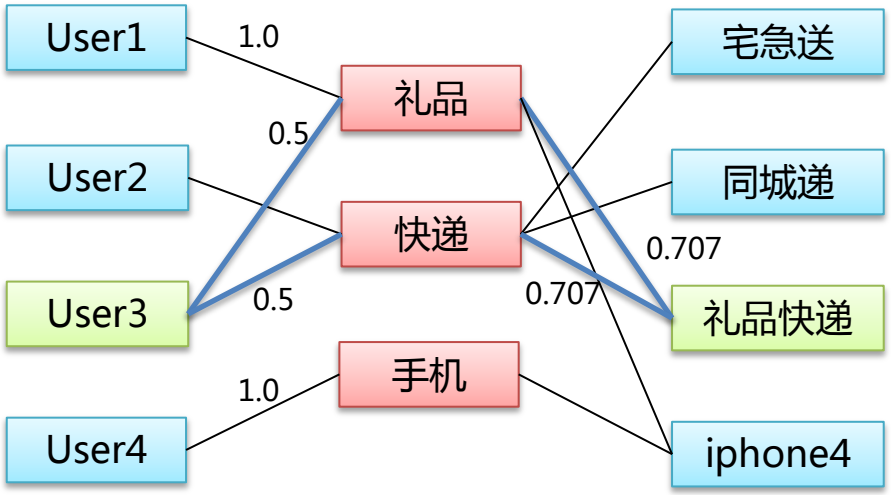
通过调用**模块Path_Method**, 计算左结点到右结点的一条路径的权重。

Path_Method输入：左右结点到某中间结点的权重
该中间结点的通过惩罚值

Path_Method输出：该**路径的权重**

Path_Method例子：路径的权重 = 左边权重 * 右边权重 * 通过惩罚值

挖掘步骤四

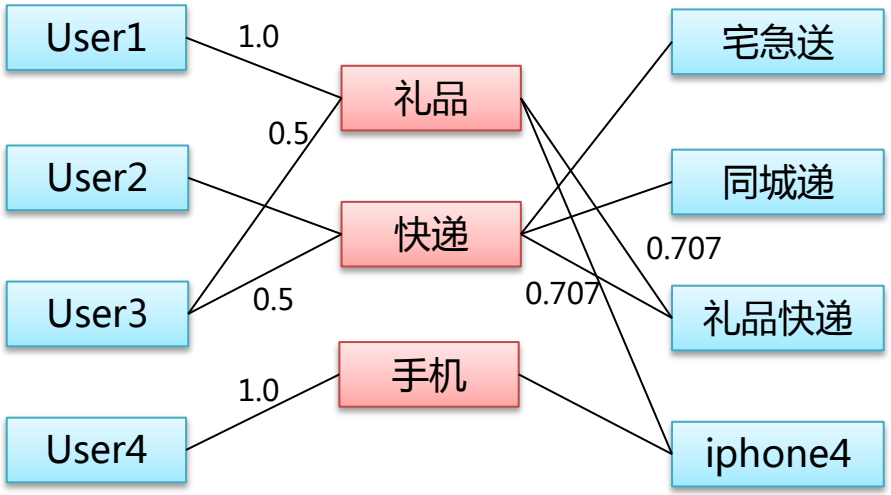


4. 相关度计算

通过调用**模块Rel_Method**, 计算左结点到右结点的相关度。

Rel_Method输入：左结点到右结点的所有路径的权重List。
Rel_Method输出：左结点到右结点的**相关度**。
Rel_Method例子：（1）相关度 = 所有权重之和
（2）相关度 = 最高的前5个权重之和/5 (Top N平均)

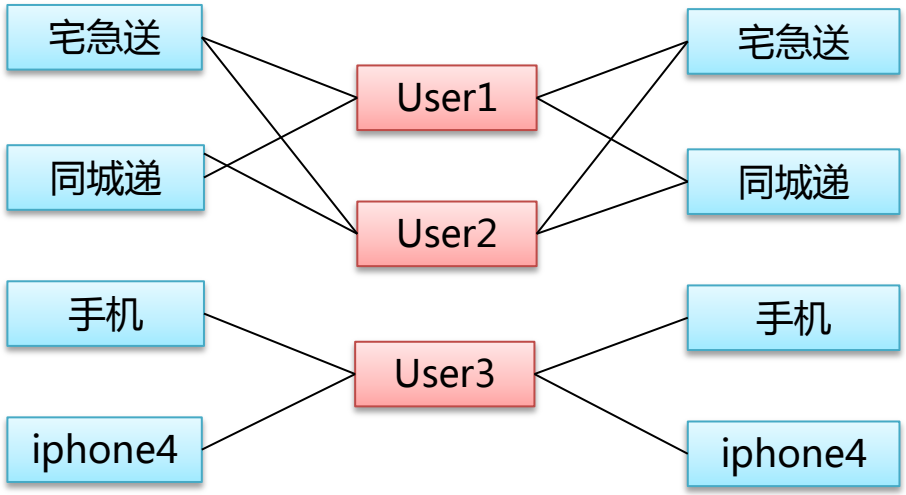
挖掘步骤总结



1. 左右结点权重归一化（模块Norm_A, Norm_B）
2. 中间结点通过惩罚（模块Punish_Method）
3. 单条路径权重计算（模块Path_Method）
4. 相关度计算（模块Rel_Method）

框架组合不同的模块（内置或自定义），实现不同的功能。

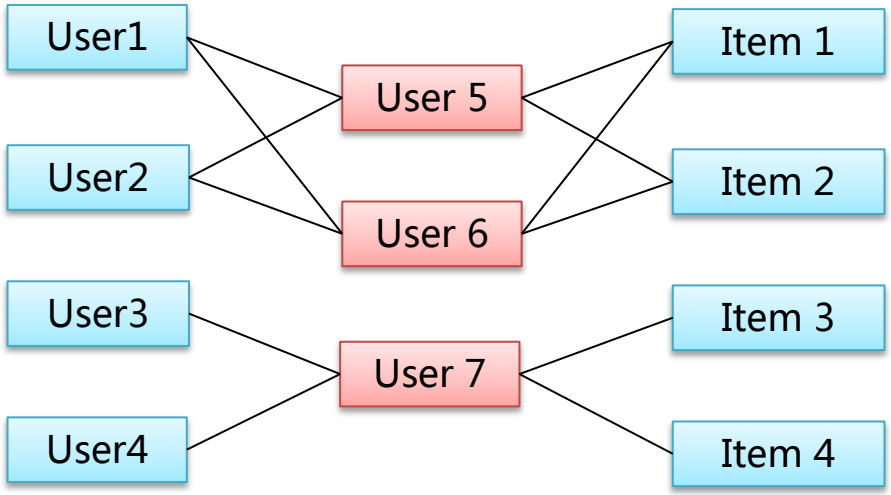
典型应用一：各种相似度指标计算



其他度量如
Jaccard系数
Dice系数
皮尔森系数
KL距离
也都能实现

	Cosine相似度	欧式距离	共现次数
边初始权重	特征权重	特征权重	全为1.0
左右结点权重归一化	L2-Norm	N/A	N/A
中间结点通过惩罚	N/A	N/A	N/A或自定义惩罚
单条路径权重计算	相乘	差的平方	相乘
相关度计算	相加	相加后取平方根	相加

典型应用二：协同过滤(以User-based为例)



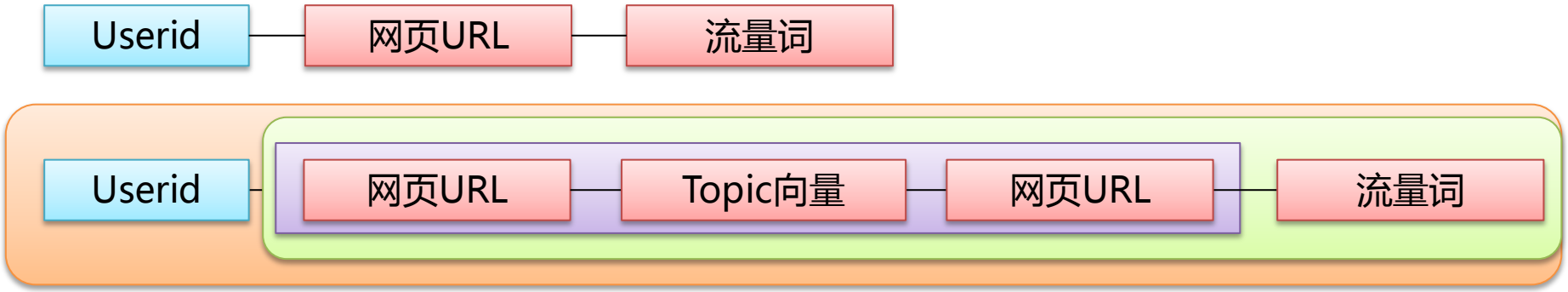
	CF
边初始权重	相似度, item权重
左右结点权重归一化	N/A
中间结点通过惩罚	N/A
单条路径权重计算	相乘
相关度计算	相加/TopN平均

典型升级过程

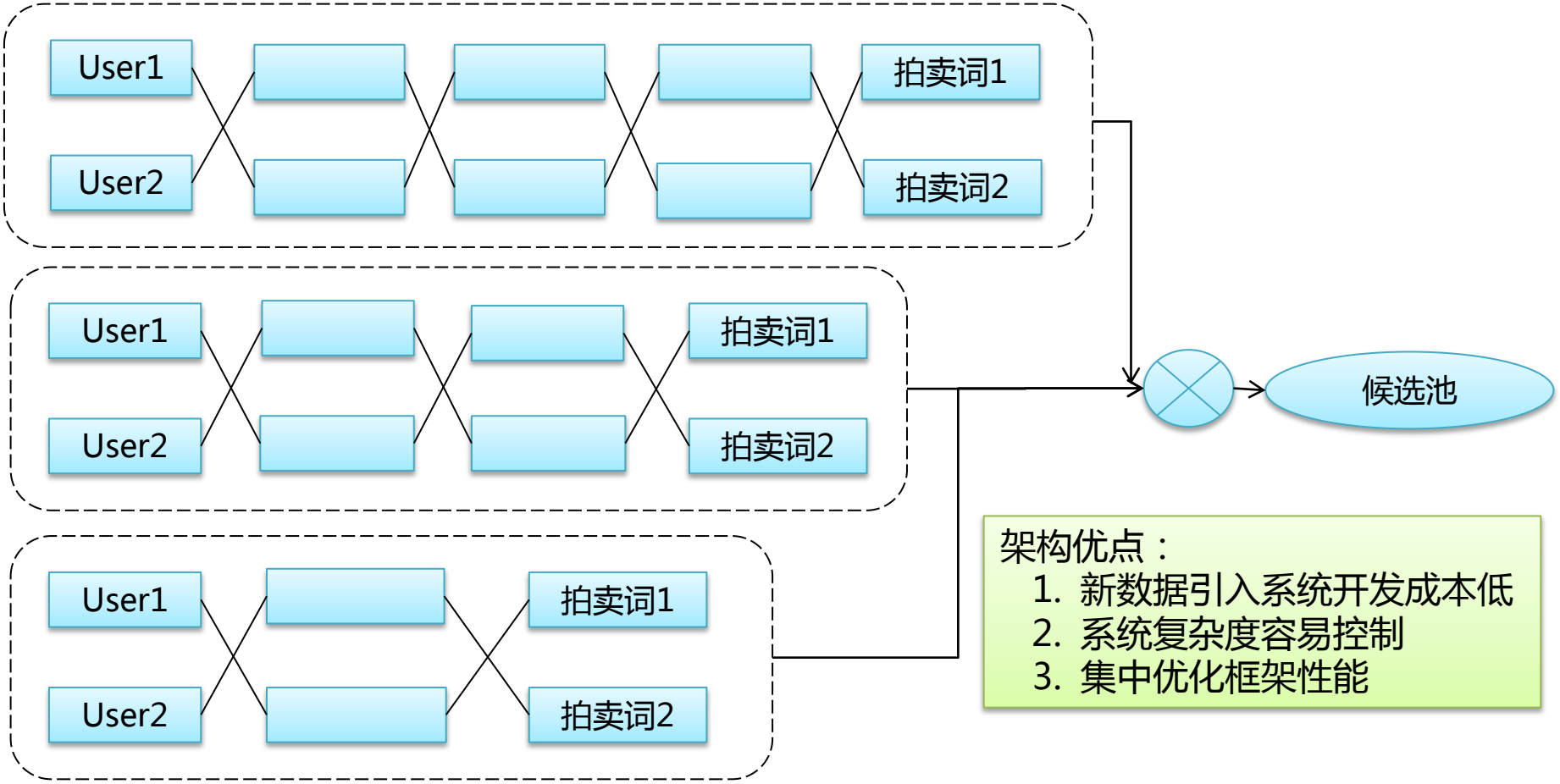
- 新数据：
  1. Userid \t 网页URL \t 流量词
     （流量词：网民搜索这些Query后会点该网页URL）
  2. 网页URL \t 网页所属的Topics向量
- Heuristics:

  1. 把广告主网站上所有网页的流量词推荐给广告主。
  2. 把广告主网站上所有网页的相似网页的流量词推荐给广告主。

本质：
(1) 发现关联关系
(2) 传递关联关系



主动推荐架构（推荐引擎部分）



Where are we?

I. 背景知识介绍
   1. 互联网广告
   2. 百度面向广告主的推荐产品
II. 设计要点
   1. 分析清楚技术目标
   2. 充分挖掘来自数据与人的信息
   3. 合理利用反馈信息
   4. 设计合适的推荐理由
   5. 注意评估结果的显著性
III. 一些经验
IV. Q&A

Baidu 百度

设计要点—合理利用反馈信息

利用反馈信息的意义：

a. 系统具有成长性。

b. 将数据积累转化成技术壁垒。

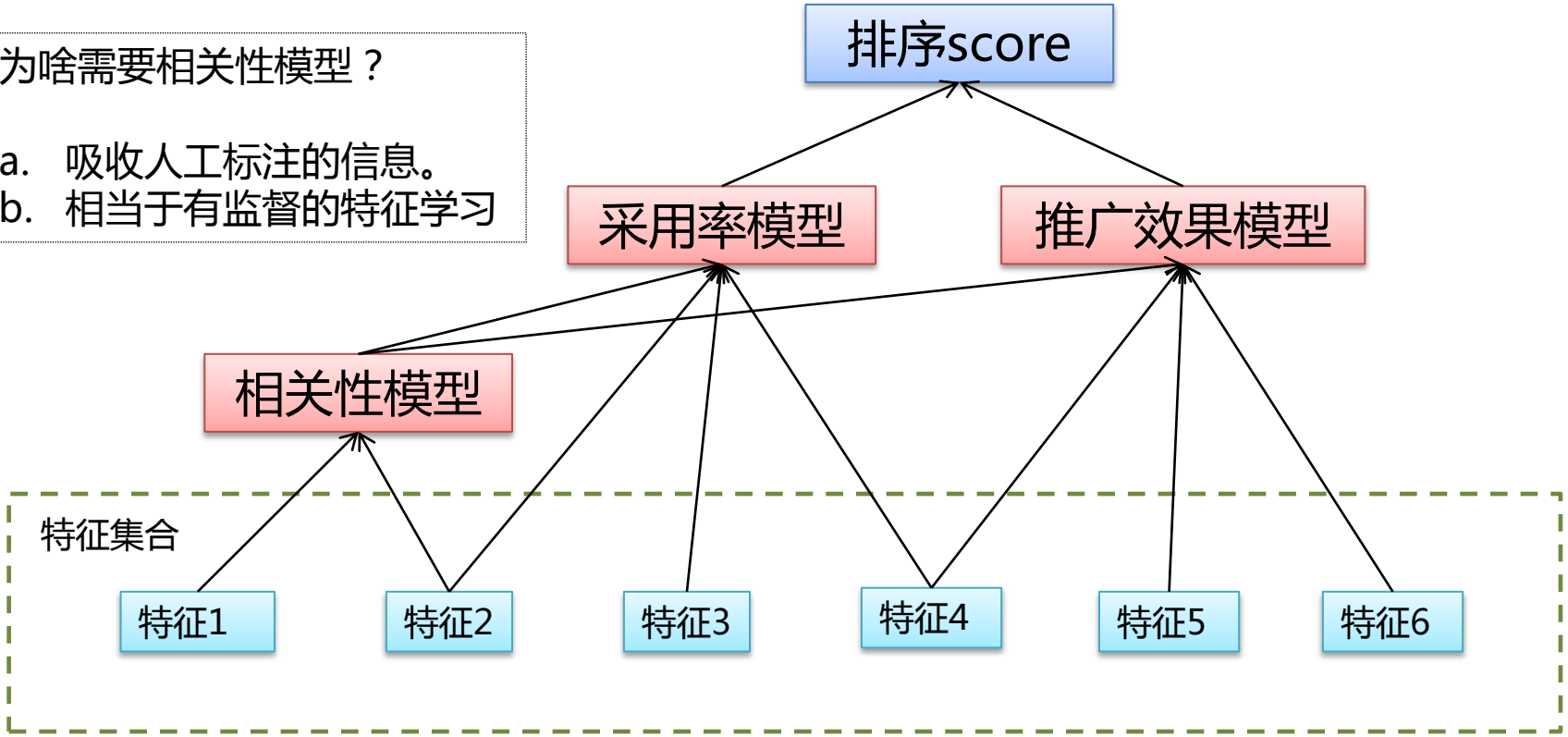
常用技术：机器学习

拍卖词推荐模型架构

P(候选词为广告主带来效果，候选词被采用)
=P(候选词被采用) * P(候选词为广告主带来效果 | 候选词被采用)

为啥需要相关性模型？

a. 吸收人工标注的信息。
b. 相当于有监督的特征学习



模型们

相关性模型：

预估值：候选与Query（及广告主）是否相关

数据：人工标注候选与Query(广告主)是否相关

模型：Adaboost

采用率模型：

预估值：P(候选词被采用)

数据：推荐结果的展现点击日志（利用了展现但未被采用的候选的信息）

模型：LR

推广效果模型

推广效果模型：

预估值：候选词为广告主带来的推广效果

数据：广告系统投放日志

模型：LR

如何衡量推广效果？

一个指标：综合广告主推广效果指标(例如：点击数，有消费率等)

数据来源：投放日志及转化跟踪数据

特征举例：

(1) 候选词被广告主购买的次数 (反映竞争激烈程度)

(2) 候选词平均展现价格与广告主平均出价的差异 (反映价格差异)

一种引入反馈信息的简单方法

if “我们没有资源构建机器学习模型”

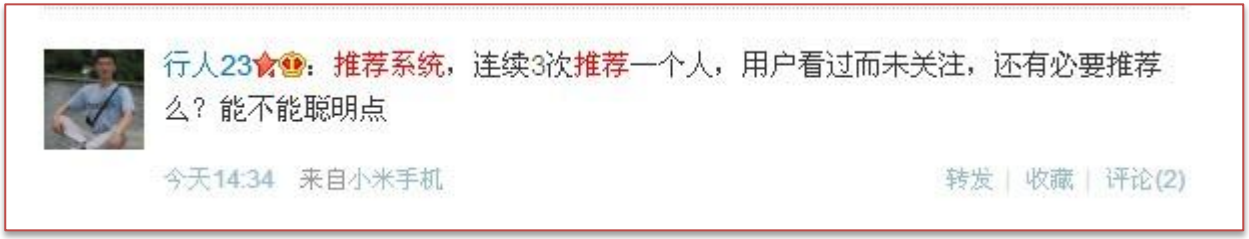
or “我们资源有限，模型的时效性跟不上” then：

简单方法：

统计每个候选被展现给每个用户但没被采用的次数，用来计算权重衰减因子, 例如：

$$S_{adjusted}(user, item) = e^{-\lambda * show(user, item)} S(user, item)$$

其中 S(user, item)为向 user 推荐 item 的原始分数，show(user, item)为最近一段时间内将 item 展现给 user 但没被采用的次数。



Where are we?

I. 背景知识介绍
   1. 互联网广告
   2. 百度面向广告主的推荐产品
II. 设计要点
   1. 分析清楚技术目标
   2. 充分挖掘来自数据与人的信息
   3. 合理利用反馈信息
   4. 设计合适的推荐理由
   5. 注意评估结果的显著性
III. 一些经验
IV. Q&A

设计要点—设计合适的推荐理由

为什么“买了X的用户中有N%还买了Y”在很多推荐系统中有效？

1. 羊群效应，从众心理。
2. 在音乐，书籍推荐中，用户无法判断Y是否自己会喜欢，只能根据对X的喜欢与否来推测Y。

为什么拍卖词推荐里没有“买了X的用户中有N%还买了Y”？

1. 竞争关系。
2. 用户可以从字面上看出自己是否会喜欢Y, 不需要根据X来推断。

面向广告主的推荐理由设计原则：

a. 尽可能多地告诉用户那些他们做决策需要的信息（例如候选词的商业指标）
b. 要说，但是“不能说太细”（防止作弊，破坏竞价机制等）

要说，但是“不能说太细”

要说，但是“不能说太细”的两种方式：

1. 模糊化具体数字（图表或区间等）
2. 将多个指标根据一定规则合成有一定意义的推荐理由。

关键词	展现理由	日均搜索量	竞争激烈程度	搜索量最高月份
苹果批发价格 HOT		30		12月
苹果手机外壳批发 HOT		50		3月
苹果手机配件批发 HOT		30		3月
苹果手机壳批发 HOT		80		3月
红苹果饰品批发 HOT		<5		-
苹果4批发 HOT		<5		-
苹果批发 NEW		30		10月
批发苹果 NEW		<5		12月
干果批发 NEW		70		12月

Where are we?

I. 背景知识介绍
   1. 互联网广告
   2. 百度面向广告主的推荐产品
II. 设计要点
   1. 分析清楚技术目标
   2. 充分挖掘来自数据与人的信息
   3. 合理利用反馈信息
   4. 设计合适的推荐理由
   5. 注意评估结果的显著性
III. 一些经验
IV. Q&A

设计要点—注意评估结果的显著性

A/B test中的显著性：

大家都知道：实验组对照组**试验后**的差异要是**显著**的（实验结束后做显著性测试保证）

容易被忽略：实验组对照组**实验前**的差异要是**不显著**的（实验前做显著性测试保证）

设计要点—注意评估结果的显著性

Problem ：

系统访问量有限，反馈数据量较小，不知道需要多长时间的实验才能得到足够的数据来得出显著的结论。

Solution：

做A/B/A测试，当A1与A2的关注指标在一段时间内都平稳后，比较A1(或A2)与B。

Where are we?

I. 背景知识介绍
   1. 互联网广告
   2. 百度面向广告主的推荐产品
II. 设计要点
   1. 分析清楚技术目标
   2. 充分挖掘来自数据与人的信息
   3. 合理利用反馈信息
   4. 设计合适的推荐理由
   5. 注意评估结果的显著性
III. 一些经验
IV. Q&A

Baidu 百度

一些经验

1. **相关性不一定是对称的**

   例：A喜欢“体育”的，B喜欢“排球”。把B喜欢的推荐给A合适，把A喜欢的推荐给B不一定合适。

2. **“修复Bug往往是最有效的策略”**

   - 算法逻辑中的bug不容易发现。需要在系统各个位置安插监控点，发现有一个不符合预期的，一定要找到合理的解释。
   - 推荐文章 “T. Raeder et. al, Design Principles of Massive, Robust Prediction Systems, KDD2012”

3. **细节决定成败**

   不要对眼里只有“算法”，把细节做好。

要点回顾

1. 分析清楚技术目标

   不要忽略推荐结果被采用以后的效用

2. 充分挖掘来自数据与人的信息

   设计可以很好融合数据与容易根据对数据的理解调整算法的系统

3. 合理利用反馈信息

   使用机器学习模型，或者一种简单的反馈系统

要点回顾（continue）

4. 设计合适的推荐理由

   尽可能多地告诉用户那些他们做决策需要的信息，尤其是商业指标，但是不能说太细

5. 注意评估结果的显著性

   保证实验组对照组实验前差异不显著，试验后差异显著

Thank You!

Question?
Comments?

Contact:

Email : jiangshen@baidu.com

新浪微博：江申_Johnson

Baidu 百度