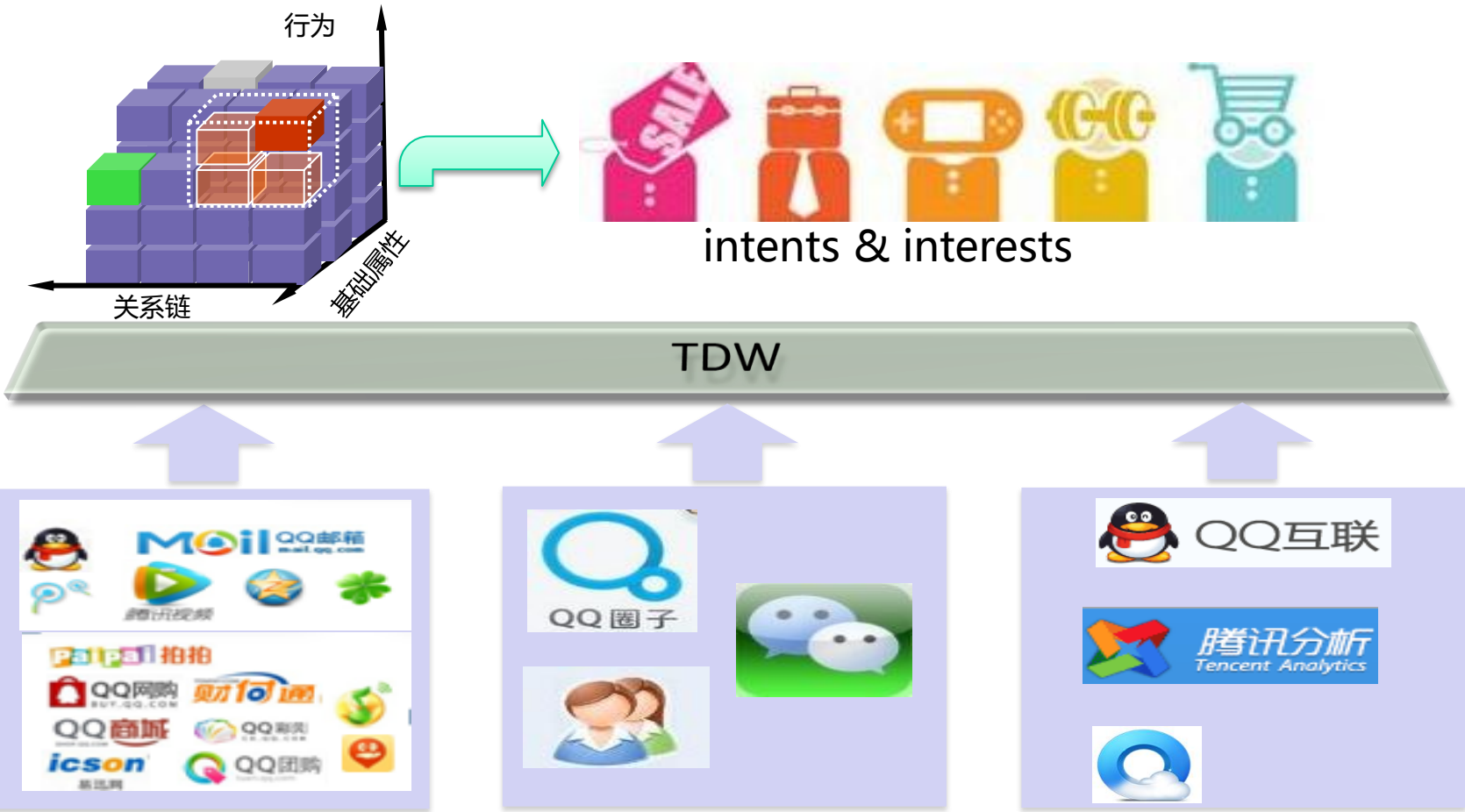
广告定向中的用户分析

广告定向组
腾讯SNG效果广告平台部

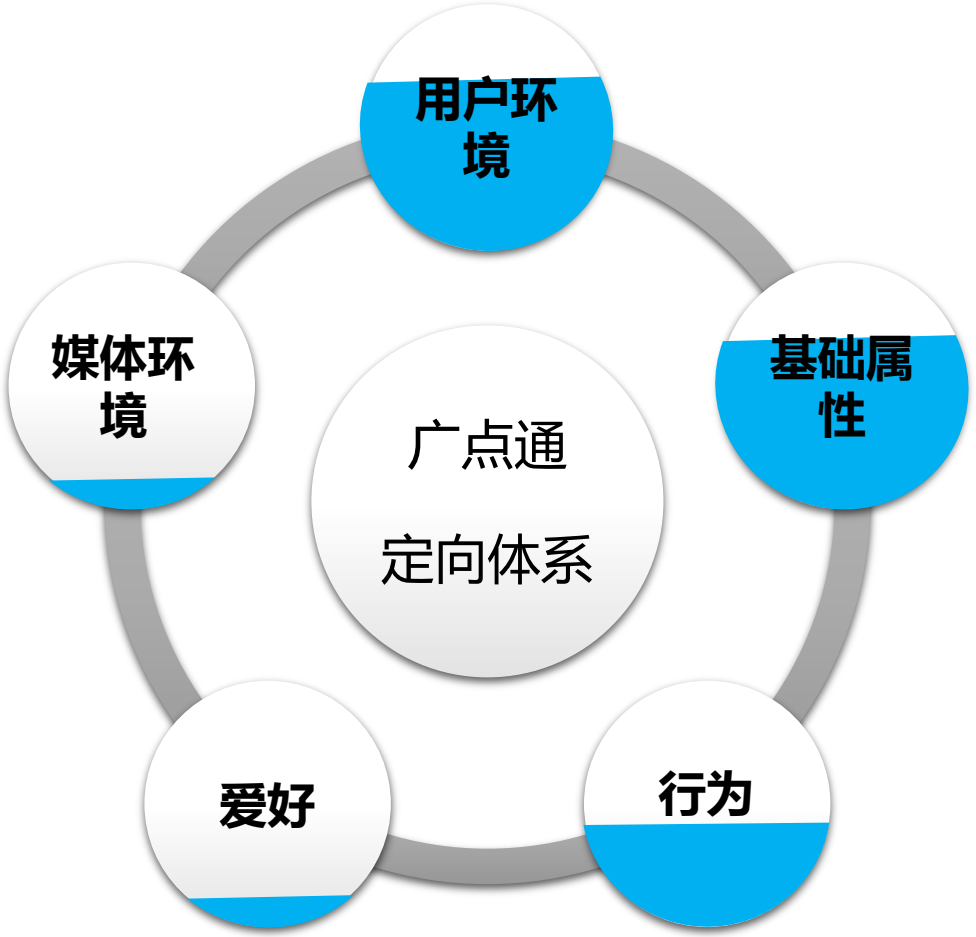
Outline

- **广点通广告定向**
- **用户分析数据挖掘工具**
  - TextMiner
  - Peacock
- **业务应用与效果**
  - 基于QQ 群的广告定向
  - 基于用户点击 URL 的挖掘与广告定向
- **用户挖掘中的问题**
  - QQ 群建模
  - 定向标签传播

腾讯用户数据挖掘


行为
关系链
基础属性
intents & interests
TDW
站内用户行为
关系链数据
站外用户行为
QQ互联
腾讯分析
Tencent Analytics
QQ圈子

广点通定向体系



广点通定向体系---用户环境

地理位置	天气状况	上网场景
• 国内 • 海外 • LBS 覆盖	• 温度、湿度、气象、紫外线 • 穿衣指数、化妆指数 • 覆盖全国所有省市	• 家庭 • 公司 • 学校 • 公共场所

广点通定向体系--- 基础属性

广点通定向体系---行为


电商购物行为
• 浏览
• 收藏
• 下单
• 成交
品牌再营销
• 到访
• 未到访
认证空间行为
• 关注
• 评论
• 转发
• 互动
PC APP行为
• 下载
• 安装
• 激活
• 活跃
• 付费
移动APP行为
• 下载
• 安装
• 激活
• 活跃
• 付费

广点通定向体系---兴趣爱好

广告定向的效果衡量

- 定向效果衡量
  - 用户覆盖量
  - 广告主使用量
  - 广告曝光量
  - 定向人群的广告点击率 (CTR)
  - 定向人群的广告转化率 (CVR)

定向方式	曝光/天	ΔCTR	ΔCVR	用户覆盖	广告覆盖
拍拍购物行为	6亿	300%	220%	1300万/月	

语义定向数据选择

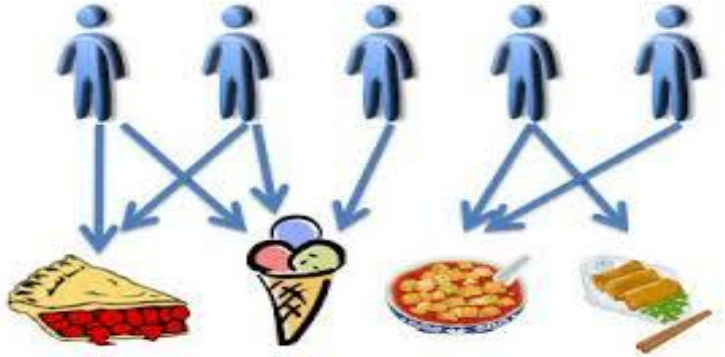
- **QQ 群数据**
  2亿 QQ 群， 覆盖 7亿用户
- **用户点击 URL**
  每天2亿点击
- **用户广告点击数据**
  每天2千万
- **电商用户行为数据**

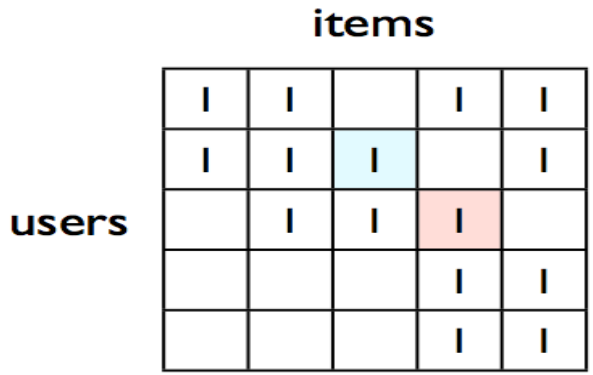
1、非隐私用户数据
2、覆盖用户多
3、商业价值高

数据挖掘

- 文本语义分析

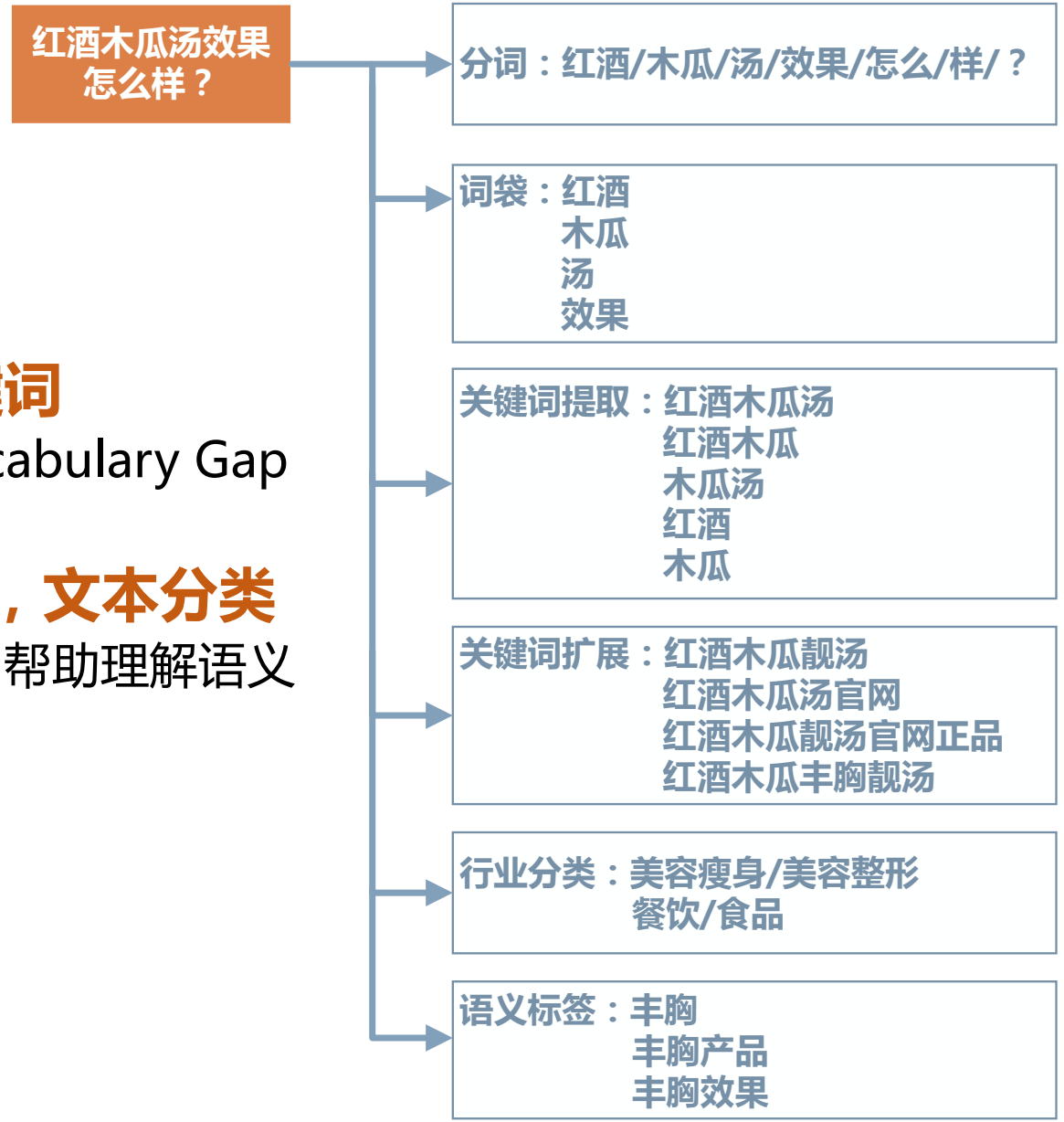
- RecSys：user-item 矩阵分解





文本语义分析

- 解决方案
  - 字面抽取：**命名实体识别、关键词**
    - 信息量小，有歧义，容易陷入 Vocabulary Gap
  - 语义分析：**文本聚类（Topic），文本分类**
    - 从海量文本数据中归纳“知识”，帮助理解语义
- 难点
  - 如何挖掘细粒度、长尾语义？



TextMiner 语义分析平台

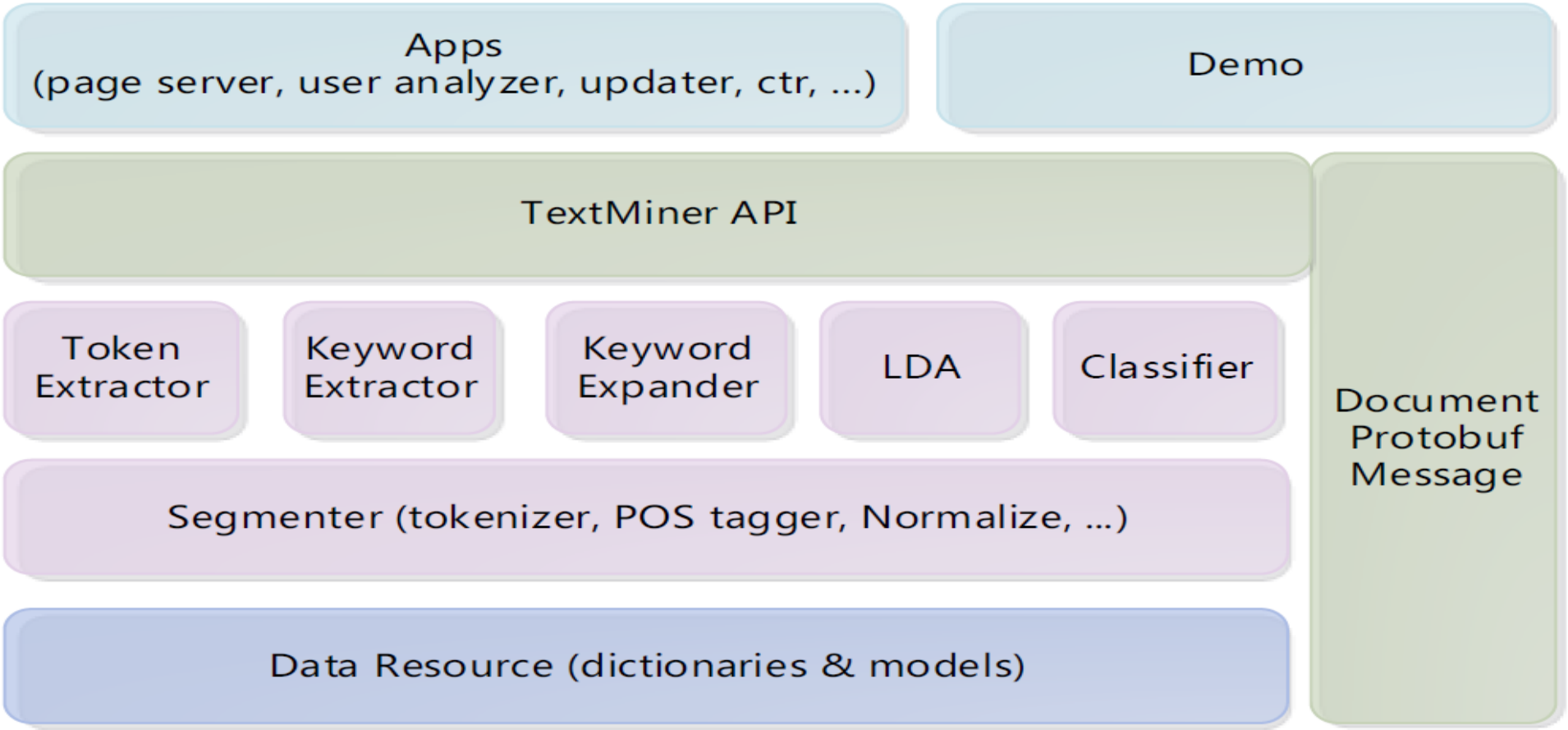
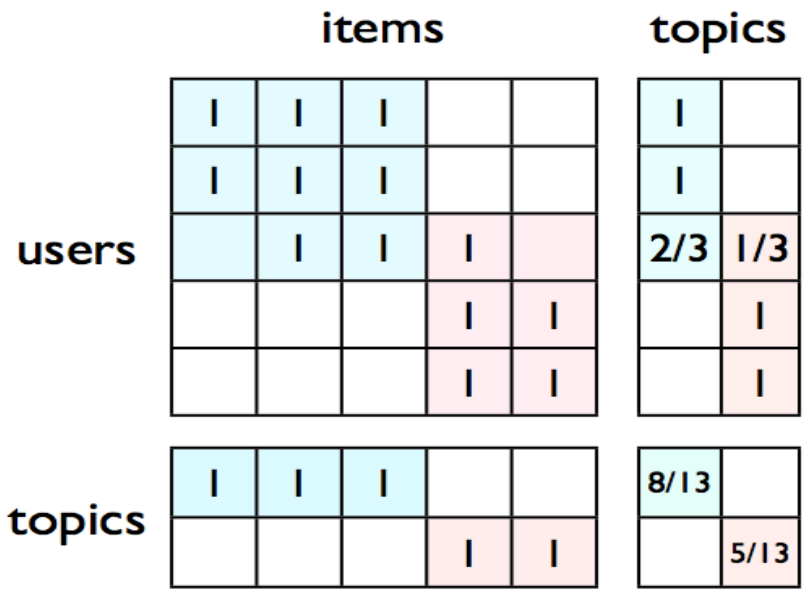
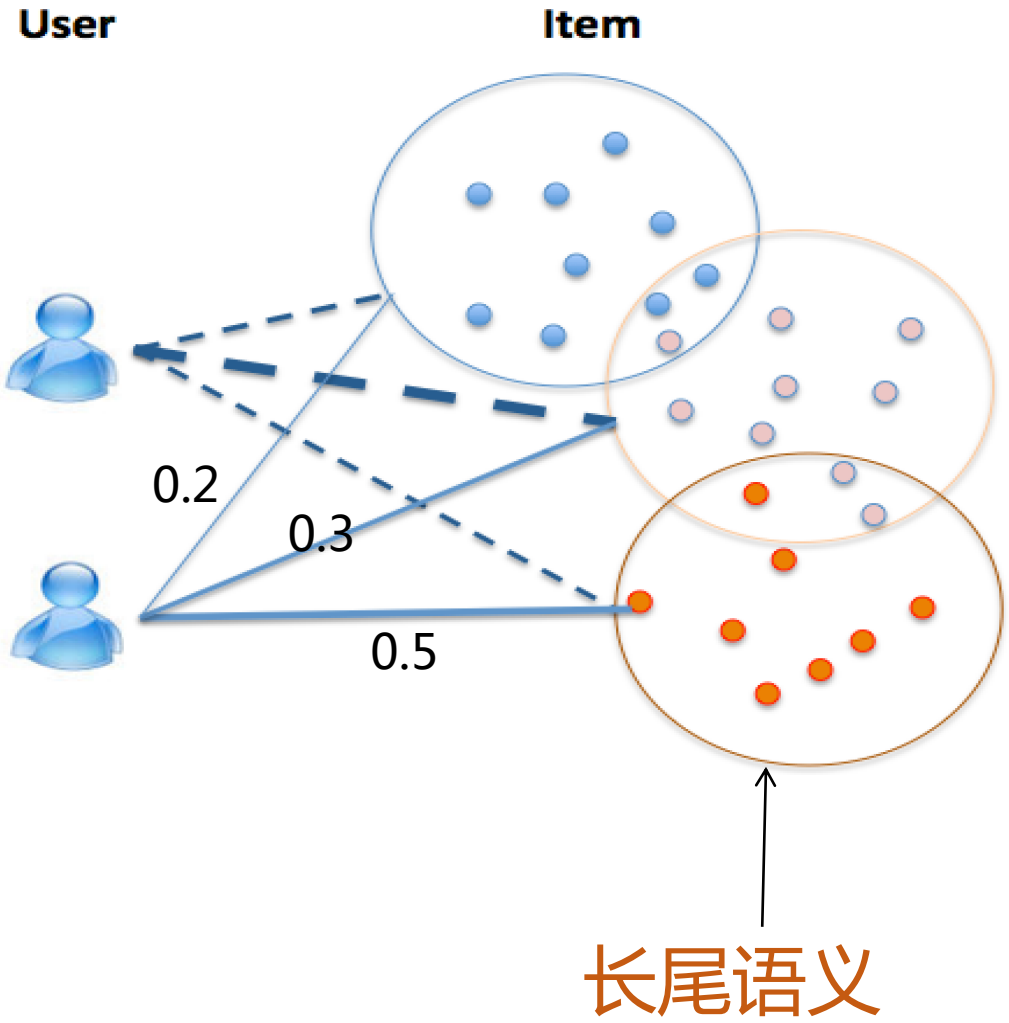


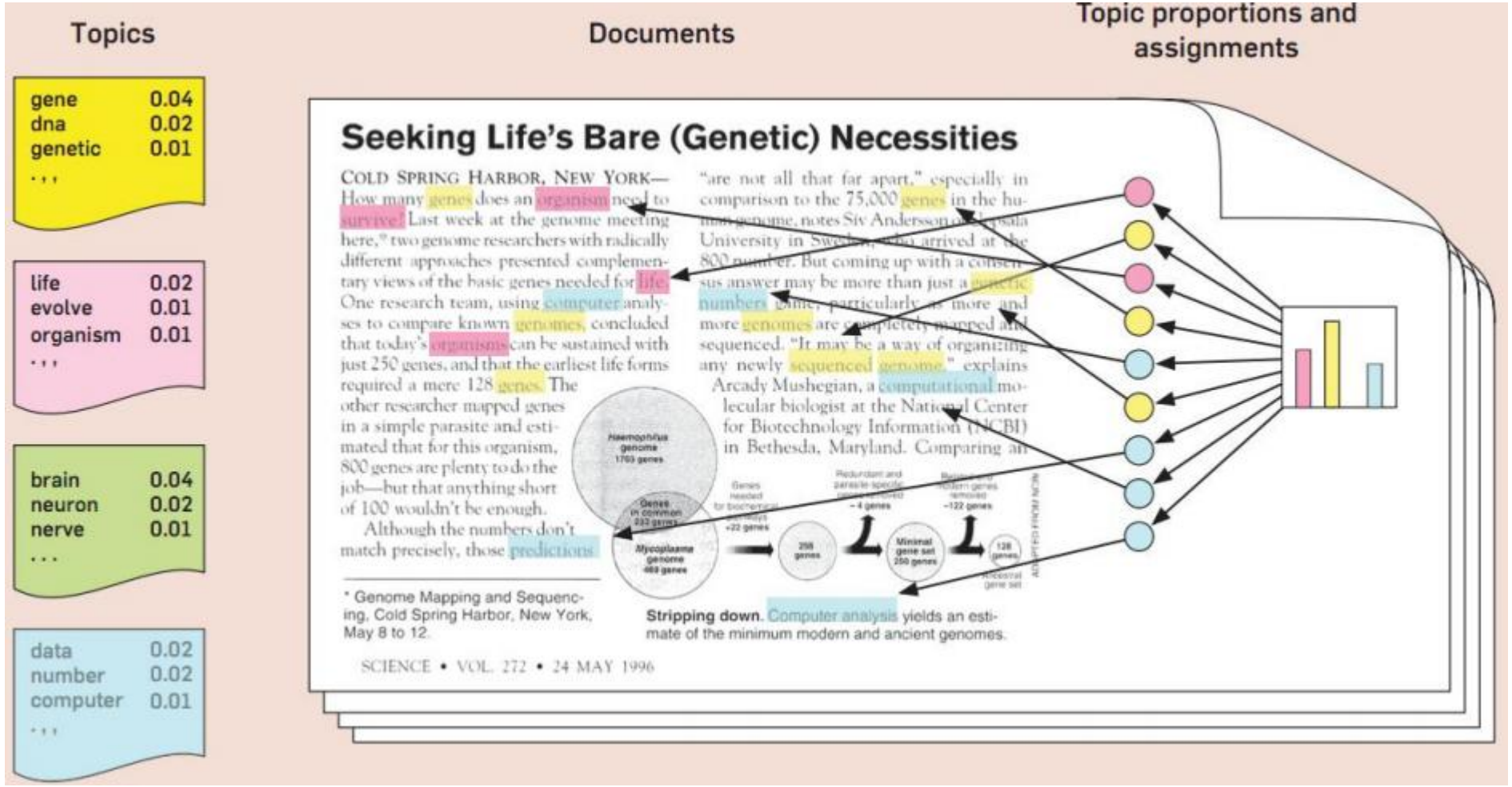
图 1 TextMiner 系统架构图

Recsys: 矩阵分解

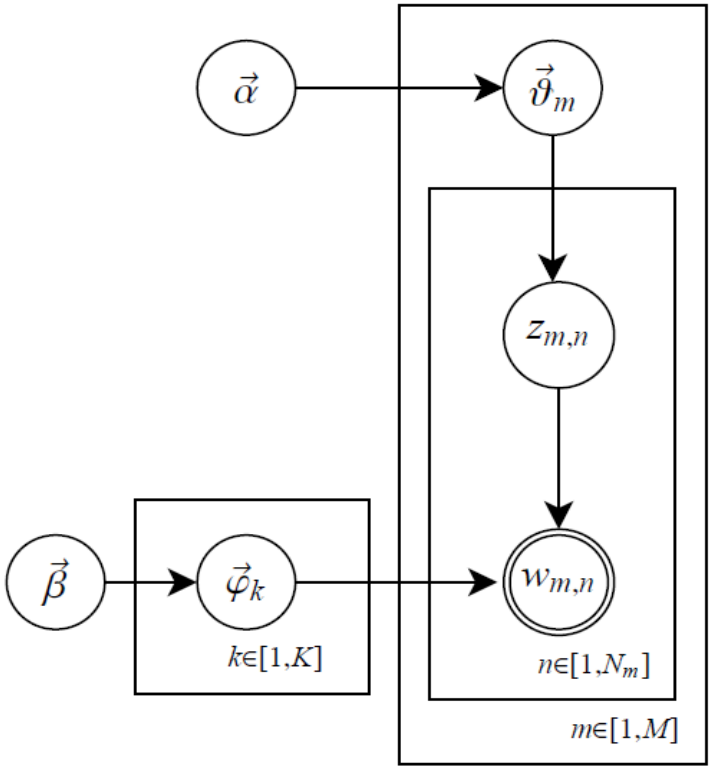

items
topics
users
topics
1
2/3
1/3
8/13
5/13

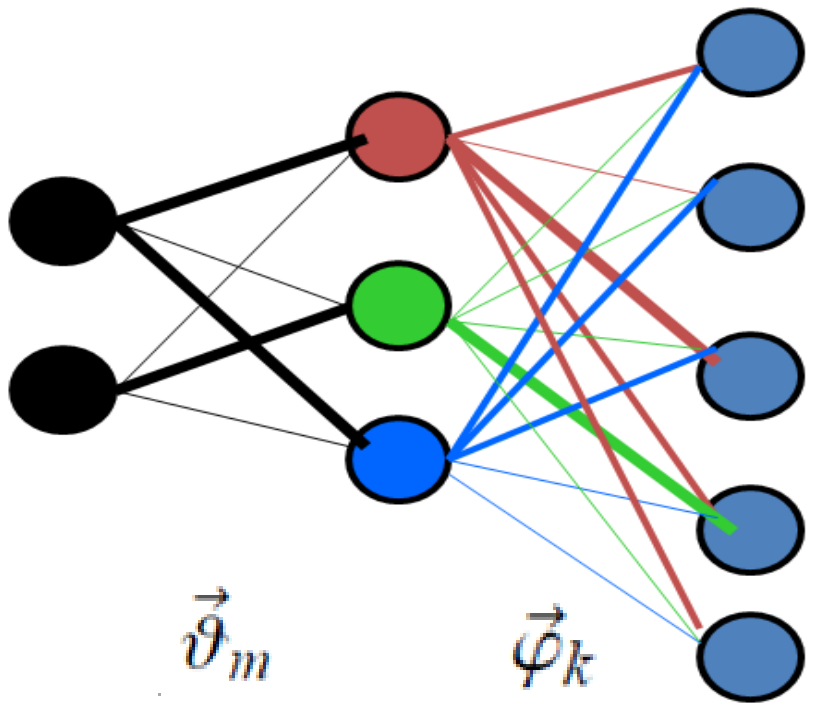

User
Item
0.2
0.3
0.5
长尾语义

LDA Topic Modeling



LDA Topic Modeling





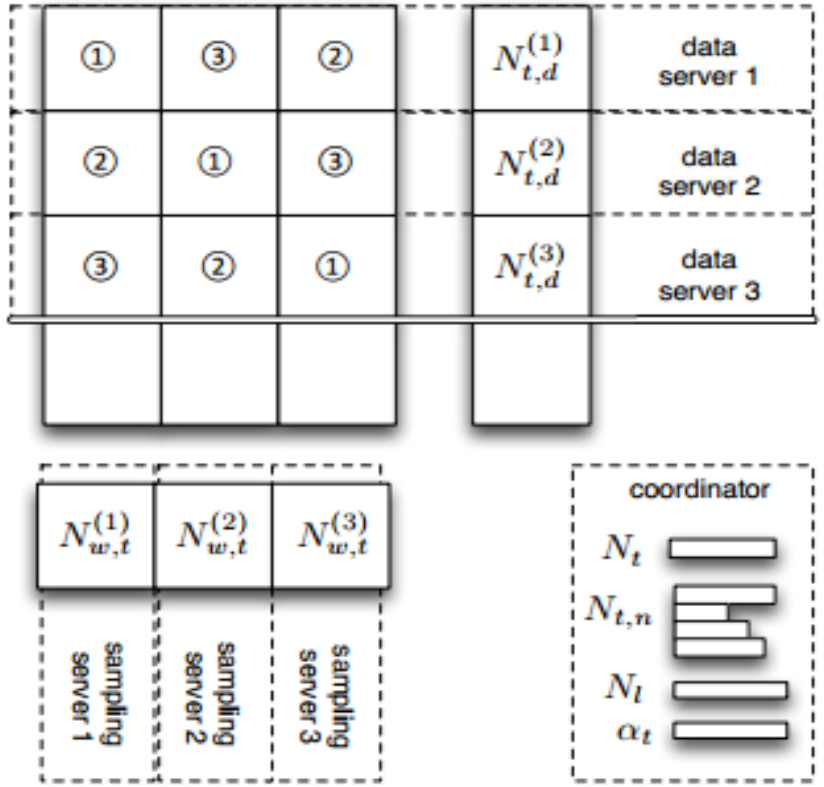
$$p(z_i{=}k|\vec{z}_{\neg i}, \vec{w}) \propto \frac{n^{(t)}_{m,\neg i} + \alpha_k}{[\sum_{k=1}^{K} n^{(k)}_m + \alpha_k] - 1} \cdot \frac{n^{(t)}_{k,\neg i} + \beta_t}{\sum_{t=1}^{V} n^{(t)}_{k,\neg i} + \beta_t}$$

P(topic|doc) **P(topic|word)**

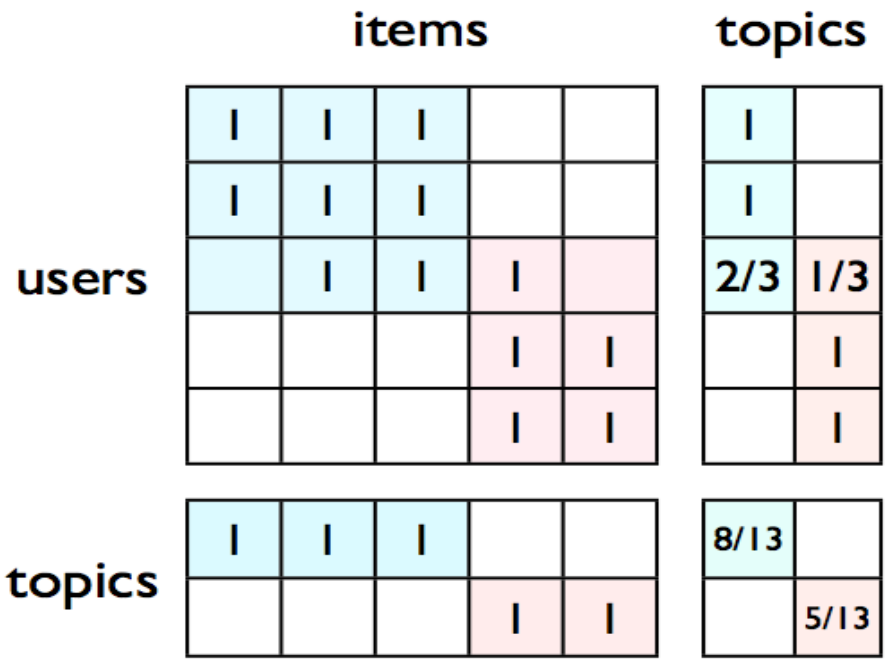
Peacock 并行矩阵分解系统

- 基于 LDA Topic Modeling 算法
- 使用Fast-LDA 算法做 Gibbs Sampling，**比标准 LDA 快30倍**
- 每轮迭代对超参数 $\vec{\alpha}$，$\vec{\beta}$ 做优化，**智能训练 topics 个数**

- 基于 Go 语言实现
- 矩阵分块并行计算
- 可以支持 **10亿 x 1亿**的矩阵分解，可以支持 **100万 topics** 计算
- 类似 Google Rephil 系统, 挖掘长尾语义,实现精准语义匹配



Peacock: 大规模矩阵分解



Matrix Type	Size
SearchQuery-word 矩阵	**10亿 x 20万**
QQ 群矩阵	**7亿 x 2亿**
QQ-URL 点击矩阵	10亿x100亿

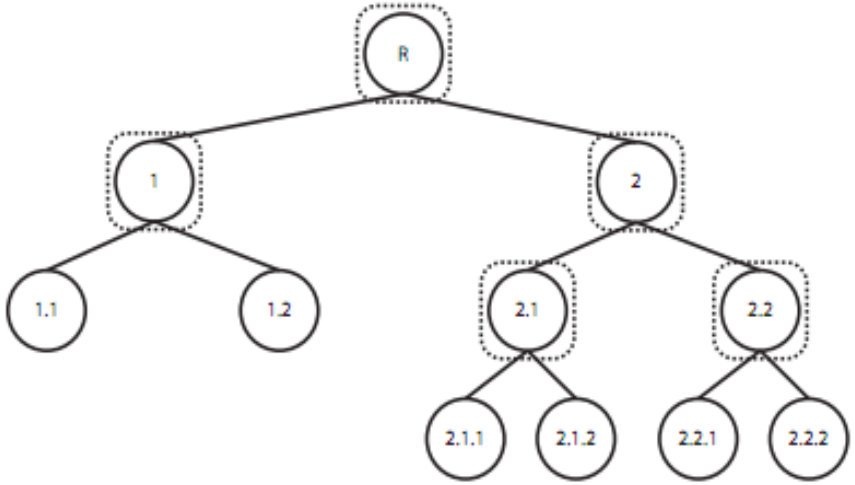
应用1：QQ群广告定向

- 群总数： 2.7亿+
- 群总用户： 7亿

数据筛选

群类别	群个数	定义
关系群	42，067，287	成员较多（成员数>20个），成员间关系相熟（关系稠密度>10%），群除了提供沟通平台，也承担着通讯录的作用
兴趣群	17，818，212	成员较多（成员数>20个），成员间关系稀疏（关系稠密度<=10%），群月度消息活跃（月消息>0），群为同好群体提供沟通平台
小团体群	37，909，781	成员较少（成员数<=20个）,消息活跃（月消息>0），功能相当于讨论组
总计	97，795，280	

QQ群层次分类器



- 圆圈表示类别节点
- 二层分类体系，一共 100+结点
- 边表示类别节点间的父子关系
- 虚线椭圆表示训练的子分类器

QQ 群广告定向

- **Peacock 模型训练**

  文本类：10 亿 query log，20w words，10w topics，160 台机器，一周训练
  关系类：5 亿 QQ，1 亿 QQ 群, 1w topics，160 台机器，2 天训练

- **分类模型训练**

  二层分类体系，一共 100+结点，MaxEnt Model　标注8万 QQ 群

- **离线效果评测**

特征集	一级行业			二级行业		
	测试样本数	准确率	召回率	测试样本数	准确率	召回率
BOW(bag of words)	12987	82.33%	80.14%	12454	79.96%	79.96%
peacock topics	12987	51.94%	57.07%	12454	39.43%	37.20%
BOW+ peacock topics	12987	**86.82%**	**84.18%**	12454	**83.05%**	79.20%
BOW+ topics + userprofile		?			?	

- **初步线上定向效果检验**

  引入5家广告主做线上 A/B test 投放测试，CTR **40%↑**

应用2：用户点击URL挖掘

- **点击URL上报**

公开信息上的URL点击 2亿/天， 来自高商业性网站的2千万/天

- **定向挖掘**

抓取用户点击URL 网页内容，对页面内容做分类
依据分类投放广告

- **线上效果**

用户覆盖提升：URL 高商业性数据每月用户6000万+，Paipai 数据用户 1200万+
显著扩充用户商业行为数据，用户覆盖提升4~5倍

定向实验效果：品牌广告主曝光提升20倍， 相对品牌广告主大盘 CTR 40% ↑
拍拍电商广告主曝光提升25%，相对拍拍广告主大盘CTR 135%↑

挖掘中的问题与难点

- **如何准确挖掘细粒度、长尾语义？**
- **高质量数据覆盖率低（按有效性排序）**
  - 商品下单，商品收藏，广告点击，搜索，搜索点击，发表，网页浏览，分享，广告浏览

问题1：QQ 群语义分析

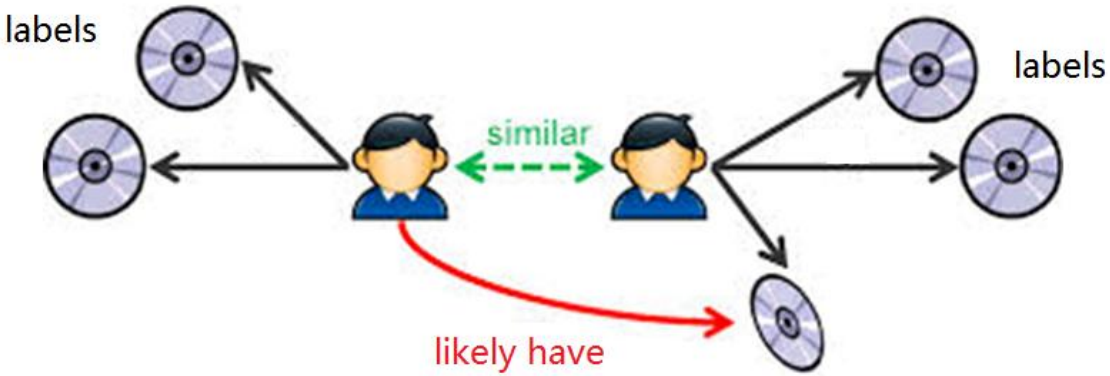
- **QQGroup 聚类的问题**
  人数上限为 1000
  不符合长尾分布
  语义相同却没有交集的人群聚类

- **QQGroup Topic Modeling**
  User-QQGroup 矩阵
  QQGroup-Word 矩阵
  融合两个矩阵到一个 model ?

问题2：高质量数据覆盖率

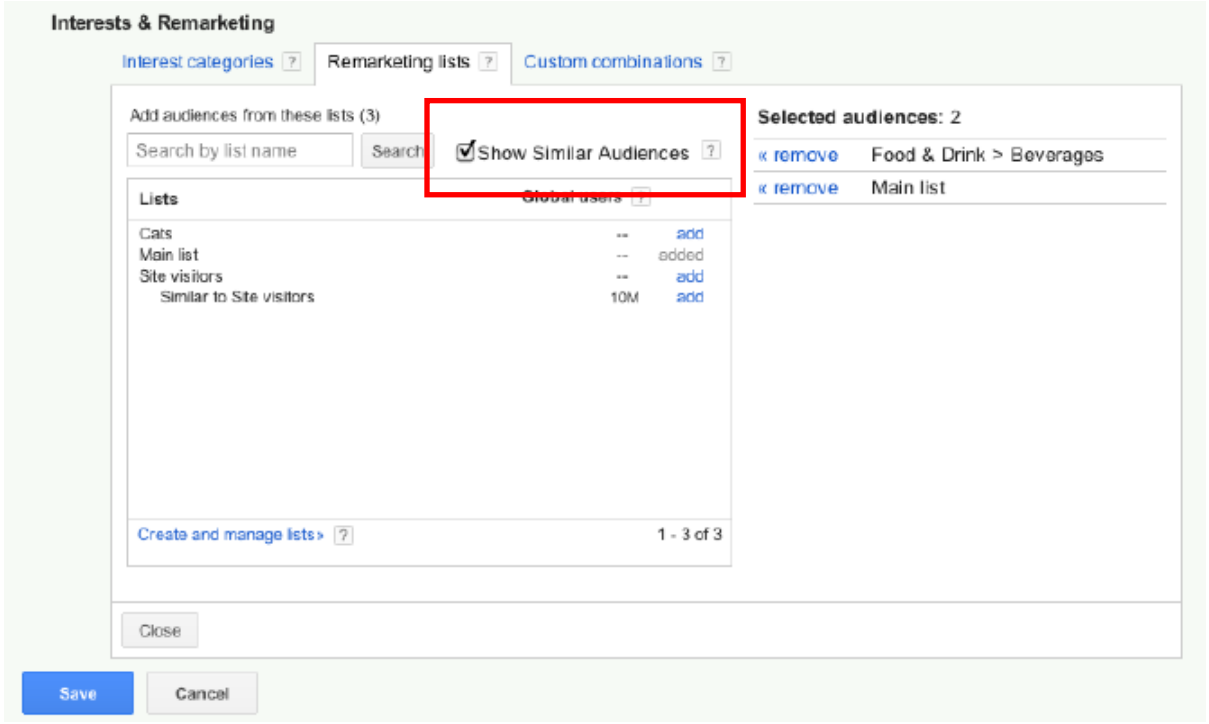
- **直接产生定向 label 的用户有限**

- **相似的的用户可能有相似的兴趣**



SimilarAudiences

AdWords looks at browsing activity on Display Network sites over the last 30 days, and uses this, along with our contextual engine, to understand the shared interests and characteristics of the people in your remarketing list.



Performance highlights:

60% More Impressions

48% More Clicks

41% More Conversions

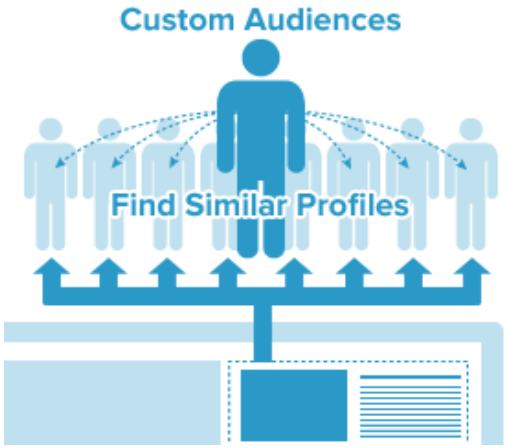
*Google Internal data 2013

https://support.google.com/adwords/answer/2676774?hl=en

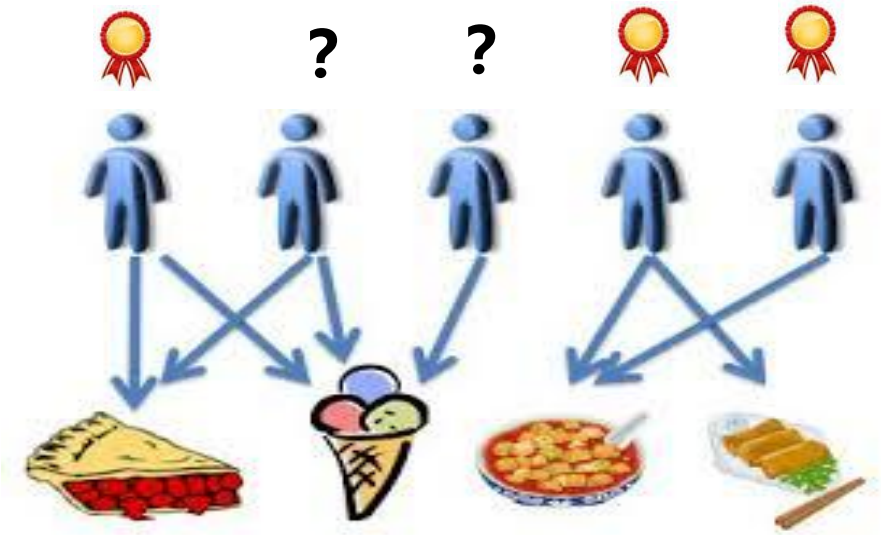
Lookalike

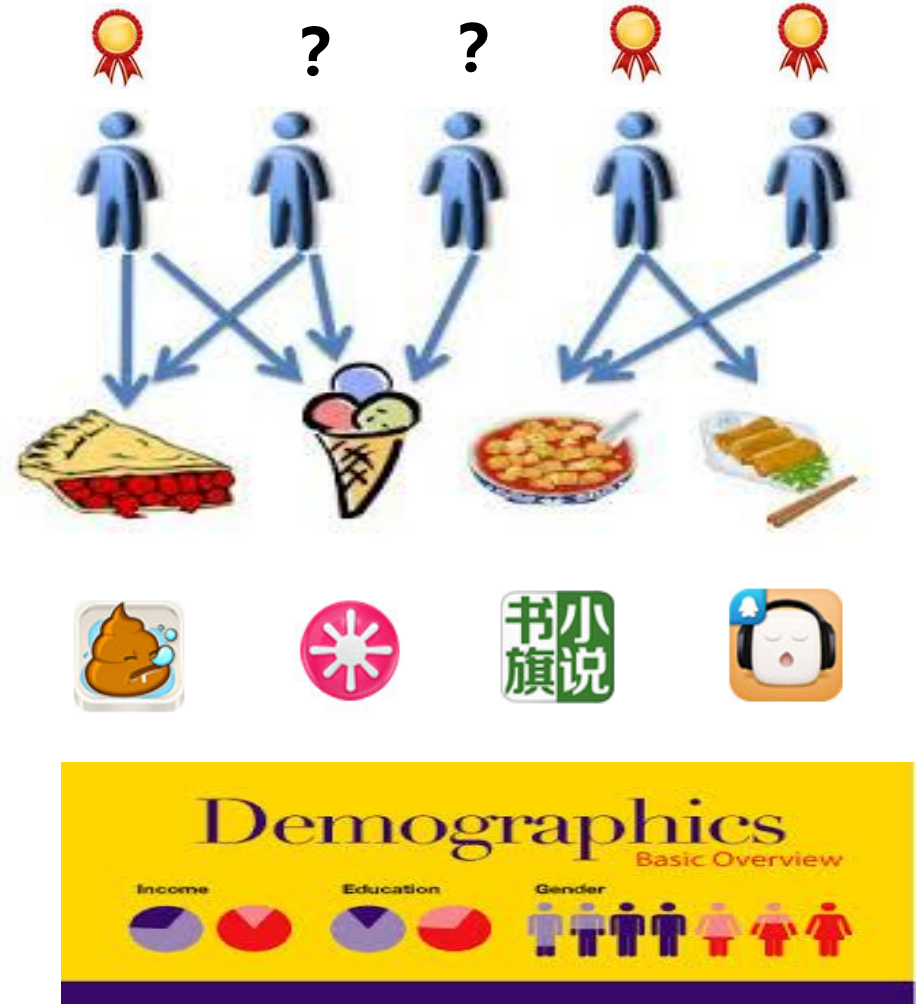
"Lookalike audiences let you reach new people who are likely to be interested in your business because they are similar to a customer list you care about."

Facebook analyzes your Custom Audience list and creates a new segment that is optimized based on either Similarity or Greater Reach.



广告定向标签


?
?


?
?
书旗小说
Demographics
Basic Overview
Income
Education
Gender

Thanks for your attentions!

扫描二维码，加入广点通

通过扫描二维码
查看岗位详情

欢迎发送简历
或联系HR