

基于短文本理解的用户画像构建

无线场景下的大数据技术实践

高君

SOGOU

摘要

- 大数据@精准营销
- 无线场景广告痛点
- 精准地域定向
- 跨域唯一用户识别
- 基于短文本理解的用户画像

大数据改变精准营销

1

多维度人群分析



2

筛选目标用户



3

目标用户量化分析



4

精准定向投放



技术体系

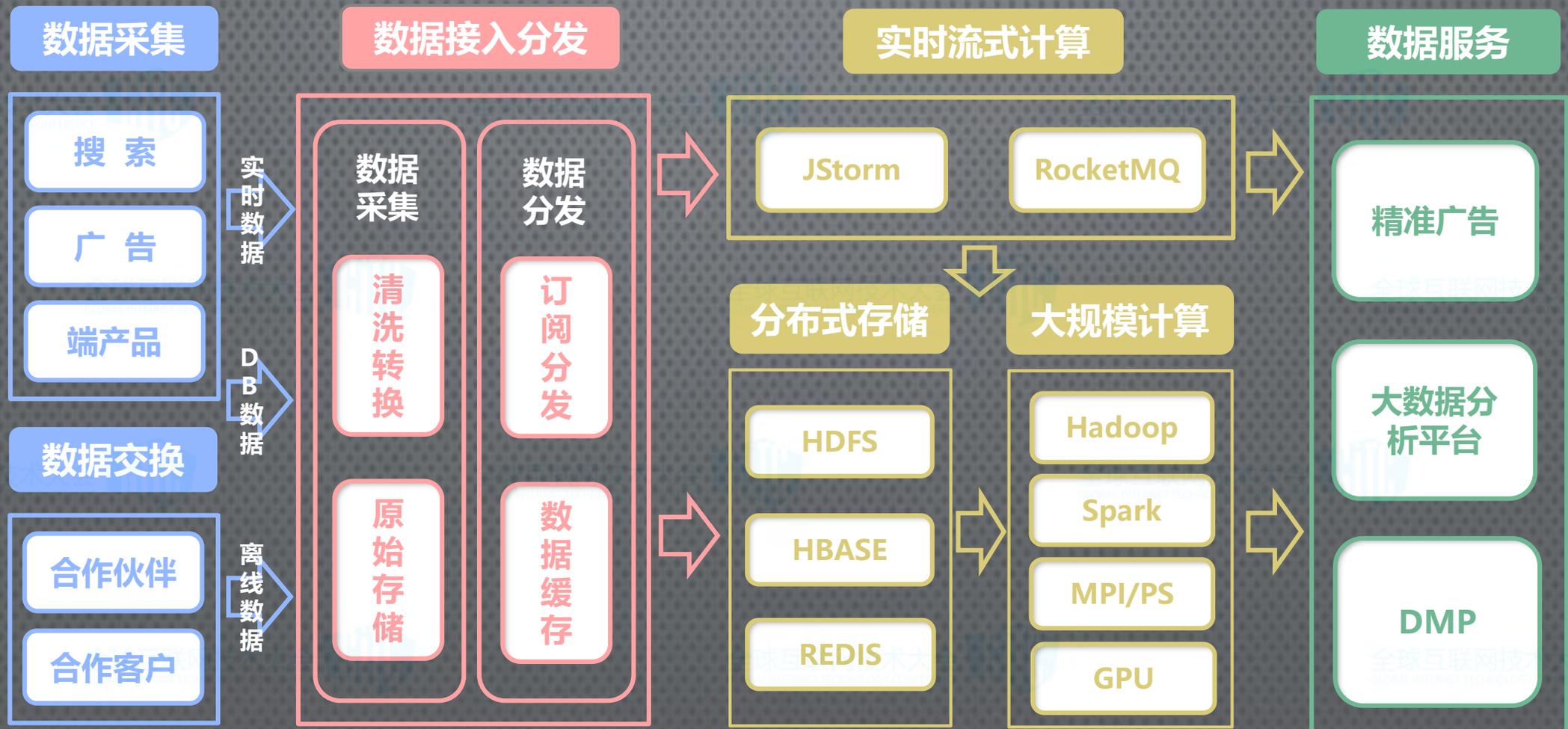
数据

认知

变现



技术体系





地域定向

- GPS
- 基站
- IP

用户识别

- PC & MOBIL
- APP & WEB

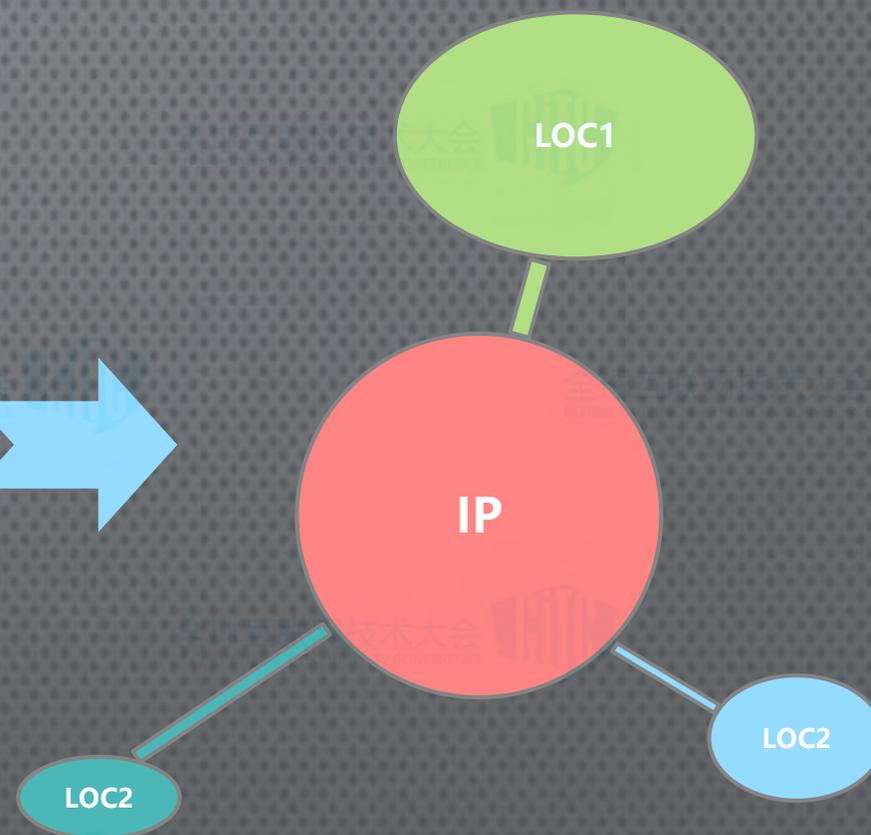
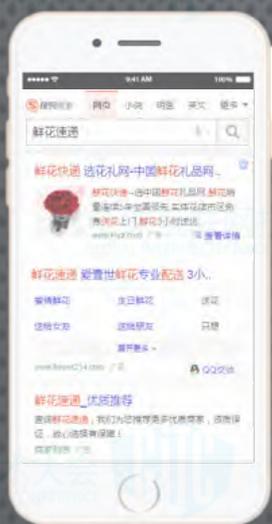
精准画像

- 人口属性
- 兴趣偏好

地域定向

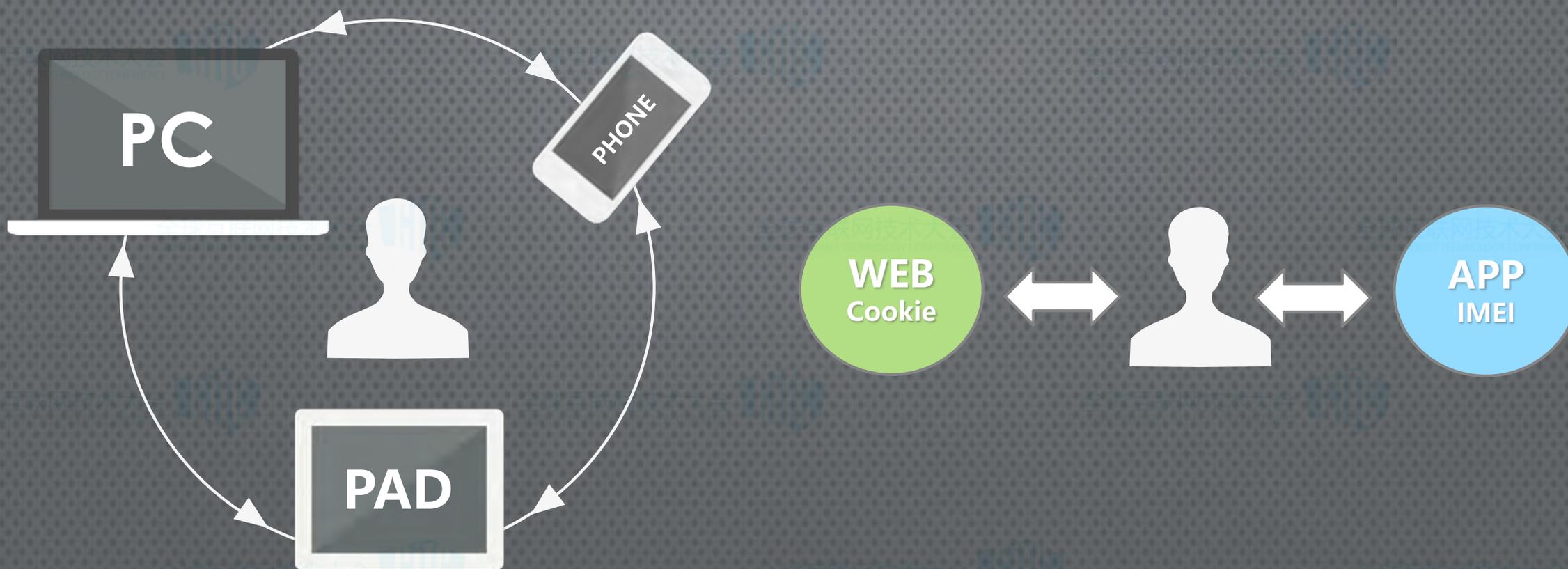


地域定向

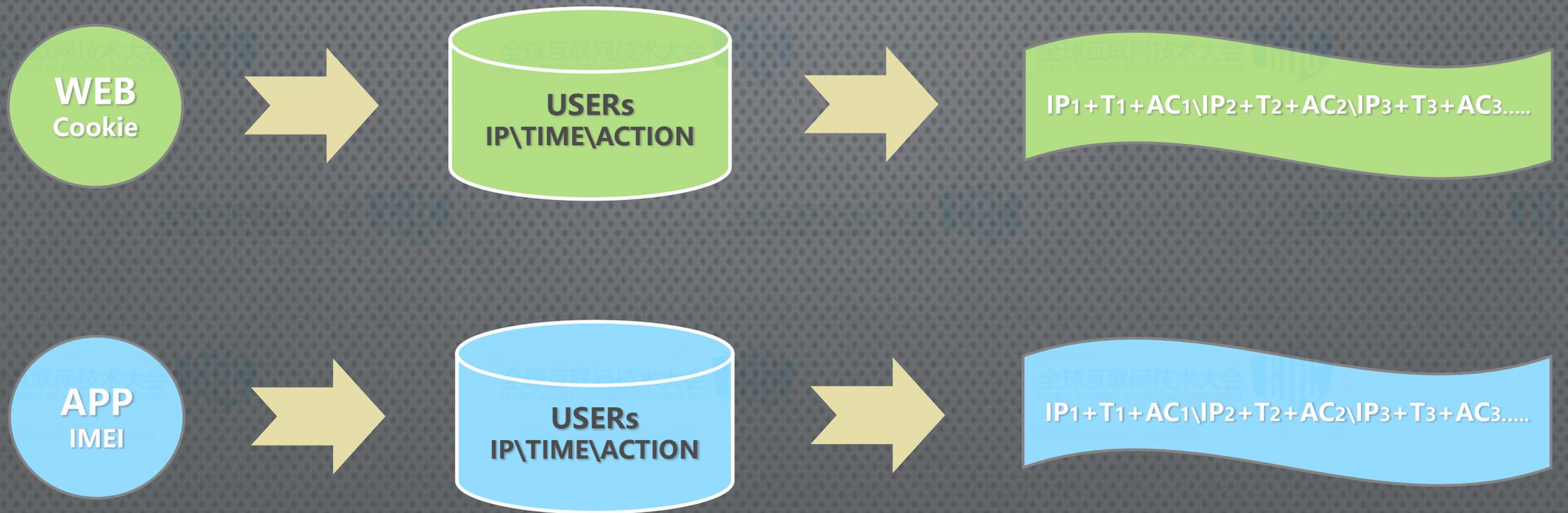


二级城市定向准确率提升50%+

用户识别



用户识别



用户识别

疑似单设备网络

IP
Time_PV

IP
Time_PV

无监督、常规相似度计算方法
准确率95%+

多设备网络

IP

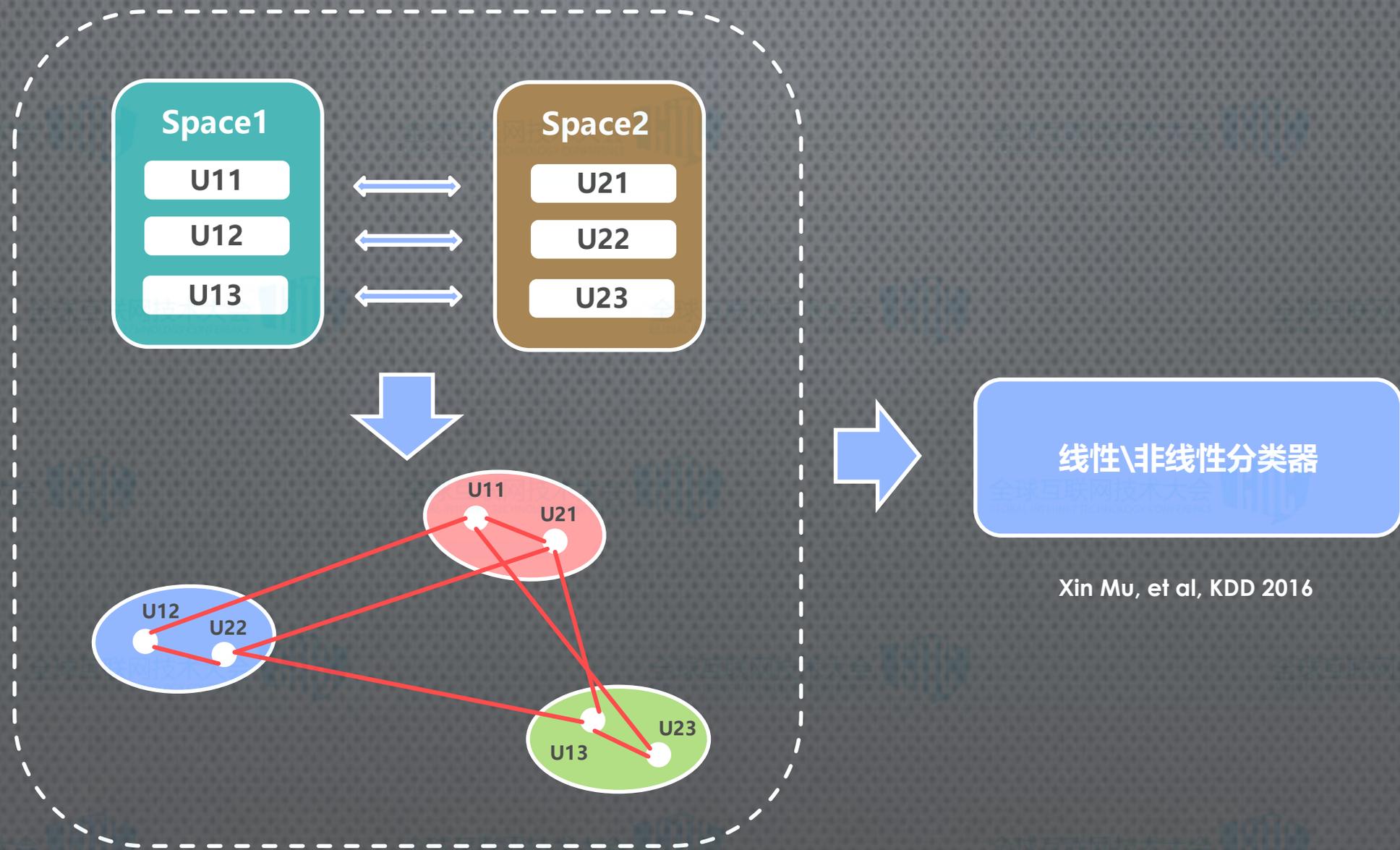
T6 : Vector
T10 : Vector
T13 : Vector
....

IP

T6 : Vector
T10 : Vector
T13 : Vector
....

有监督、常规分类算法(LR\GBDT)
准确率90%+ , 召回率提升20%

用户识别



Xin Mu, et al, KDD 2016

精准画像

一级标签

人群属性

兴趣爱好

媒体偏好

消费能力

二级标签

基本信息

地域

行业兴趣

二级行业兴趣

品类兴趣

媒体行业偏好

媒体访问时长

触点偏好

购物偏好

购买力

消费等级

三级标签

性别
学历
...

年龄
常住城市

金融
在线旅游
理财
...

汽车
新闻
二手车

阅读
媒体浏览时长
广告信任度
...

新闻
常用搜索
PC/无线

电商网站偏好
6次以上购买
同类商品点击率
...

品牌偏好
消费等级

消费者A，女，25岁左右，住在北京，常去广州、深圳，金融行业，喜欢美剧、小说、网上购物，常在iPhone手机上网，喜欢购买短靴、风衣等品类，是X品牌忠实粉丝，一段时间内曾购买N次。受促销和广告影响大。常去购物网站、金融网站和时尚网站。度假意愿为海岛，广告信任度高。

精准画像



厨宝烤箱、世情薄,人情恶,雨送黄昏花易落,晓风干,泪痕、处女座代表的花朵、烤鸡胸肉的做法、烤箱、联塑pvc排水管规格表、大王椰、性格文静是什么意思、250ml牛奶用多少克奶粉冲、化蝶去寻花 夜夜栖芳草什么意思、福睿斯、斗鱼tv、厨宝是什么、发酵箱



通用画像：长期兴趣、短期需求

- 多层级、全行业

垂直行业画像：目标导向

- 重效果、垂直行业

短文本理解

文字
歧义大

文本
过短

覆盖率和准确
率权衡

短文本预处理

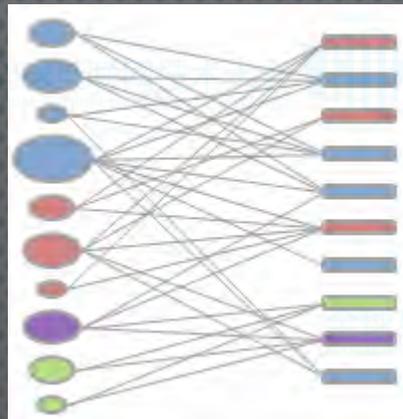
短文本扩展



1. Issue x as a query to a search engine S .
2. Let $R(x)$ be the set of (at most) n retrieved documents d_1, d_2, \dots, d_n .
3. Compute the TFIDF term vector v_i for each document $d_i \in R(x)$.
4. Truncate each vector v_i to include its m highest weighted terms.
5. Let $C(x)$ be the centroid of the L_2 normalized vectors v_i :
$$C(x) = \frac{1}{n} \sum_{i=1}^n \frac{v_i}{\|v_i\|_2}$$
6. Let $QE(x)$ be the L_2 normalization of the centroid $C(x)$:
$$QE(x) = \frac{C(x)}{\|C(x)\|_2}$$

Mehran Sahami et al, WWW 2006

基于搜索点击数据的SimRank



SimRank:

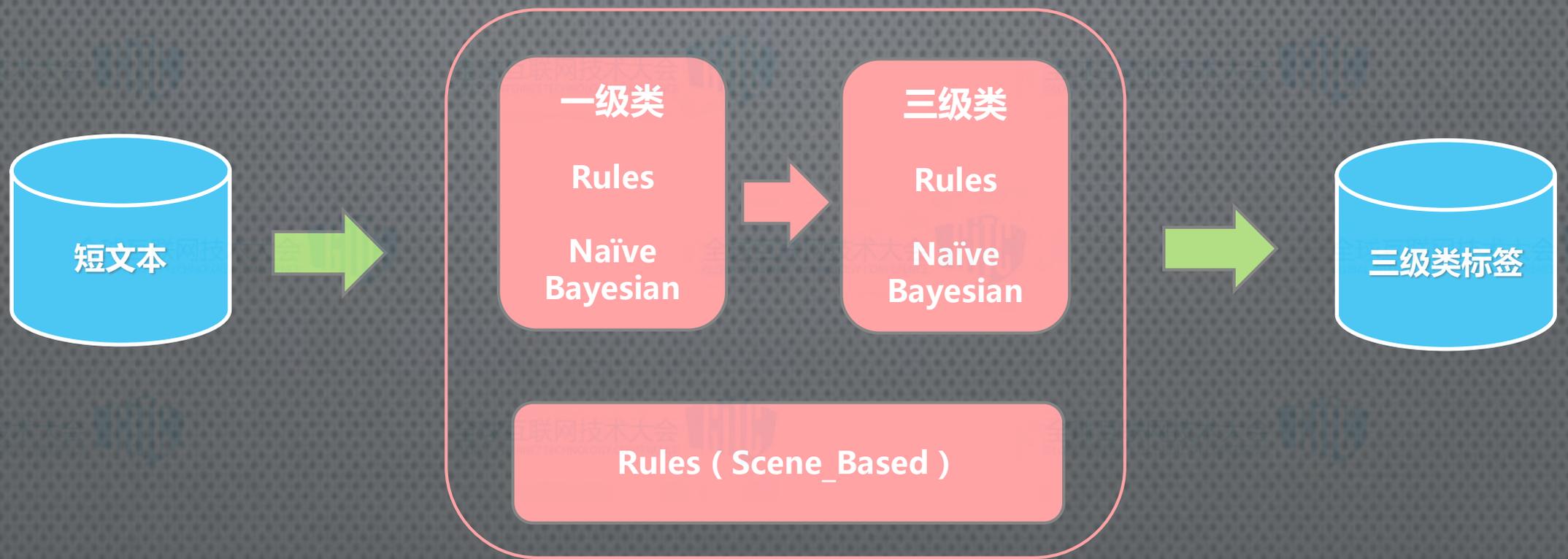
$$Sim_k(q, q') = \frac{C_1}{|E(q)||E(q')|} \sum_{l_u \in E(q)} \sum_{l_u \in E(q')} Sim_{k-1}(l_u, l_u)$$

$$Sim_k(u, u') = \frac{C_1}{|E(u)||E(u')|} \sum_{l_q \in E(u)} \sum_{l_q \in E(u')} Sim_{k-1}(l_q, l_q)$$

Query-URL Bipartite-Graph

Ioannis Antonellis, et al, WWW 2004

通用画像：2000+类



层级递进分类；基于规则的分类策略 + 基于Bag of Words的朴素贝叶斯分类器；准确率90%+，召回率80%+

通用画像：2000+类

- N-GRAM (修复NAÏVE假设的误差) : 1GRAM , 2GRAM , 3GRAM

- NAÏVE BAYES

- 先验概率 : $p(c_j) = D_{c_j} / \sum D$

- 条件概率 : $p(w_i | c_j) = \text{count}(w_i, c_j) / \sum \text{count}(W, c_j)$

- 后验概率 : $p(c_j | W) = p(c_j) \cdot p(W | c_j) / p(W) = p(c_j) \cdot \prod_i p(w_i | c_j) / p(W)$

- 优化

- LAPLACE平滑 :

- $p(w_i | c_j) = [\text{count}(w_i, c_j) + 1] / [\sum \text{count}(W, c_j) + |V|]$

- FEATURE选择 : 信息熵 (LAST 1‰ , 10W个FEATURE)

- $H(\text{feature}_i) = -\sum_{j=1}^n p(c_j) \log p(c_j)$

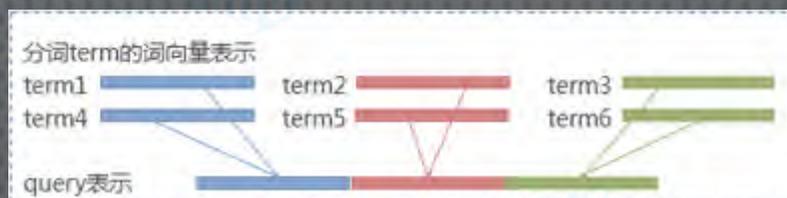
- 场景修正 :

- $p(c_j | W) = \text{bias} * p(c_j | W)_{ori}$

- 置信度 : 判断分类器的区分度 , 准确率 86% -> 93%

垂直画像：电商 金融

- 基于word-embedding的短文本向量化表示
 - Term向量的生成
专有语料库：eg. 查询词对应的商品标题集合
 - Term向量聚类：
K-means，以内积相似度作为聚类指标，类簇数目：128
 - 短文本向量化表示
基于词向量聚类的average-pooling的融合方式
查询词特征维度：D*K（D为词向量特征维度，K为词向量聚类数目）



- 短文本分类 Softmax Regression \ SVM (One vs Others)

层级递进分类；基于Word-Embedding的线性\非线性分类器；准确率较NB提升15%+，召回率提升30%+

搜狗大数据

洞悉看不到的商业秘密



用户登录

请输入用户名

请输入密码

请输入验证码

登录

忘记密码

多维度分析报告--精准、价值、易读



市场格局研究
趋势分析, 品牌认知



消费者属性分析
识别消费者



消费者偏好分析
洞悉消费者



规划营销方案
触及消费者

