
51CTO
WOT 2015
“互联网+”时代
大数据技术峰会
中国·深圳 | 2015.11.28-29
解码数据未来

社交数据在征信领域的应用探索

刘黎春　　SNG, 腾讯

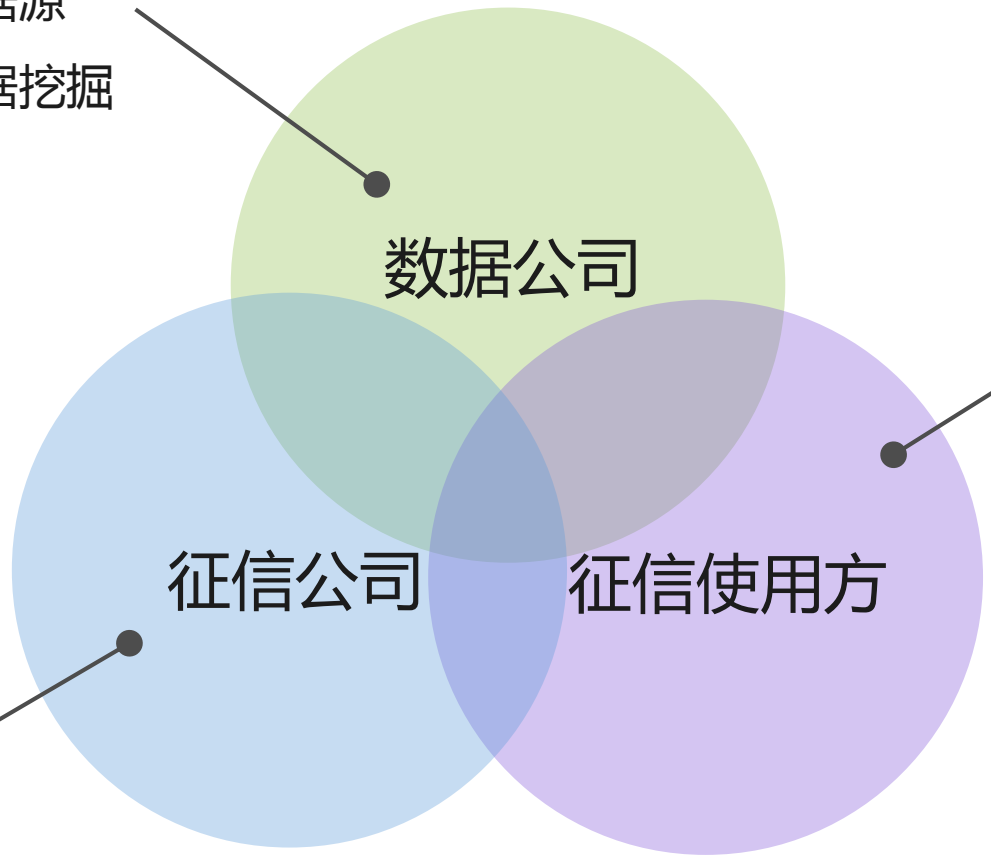
- 社交征信背景
- 腾讯社交网络数据
- 个体用户画像研究
- 社团圈子研究
- 模型建设及应用

社交征信背景

传统征信相关机构

数据的采集和提供

- 独特的数据源
- 初步的数据挖掘



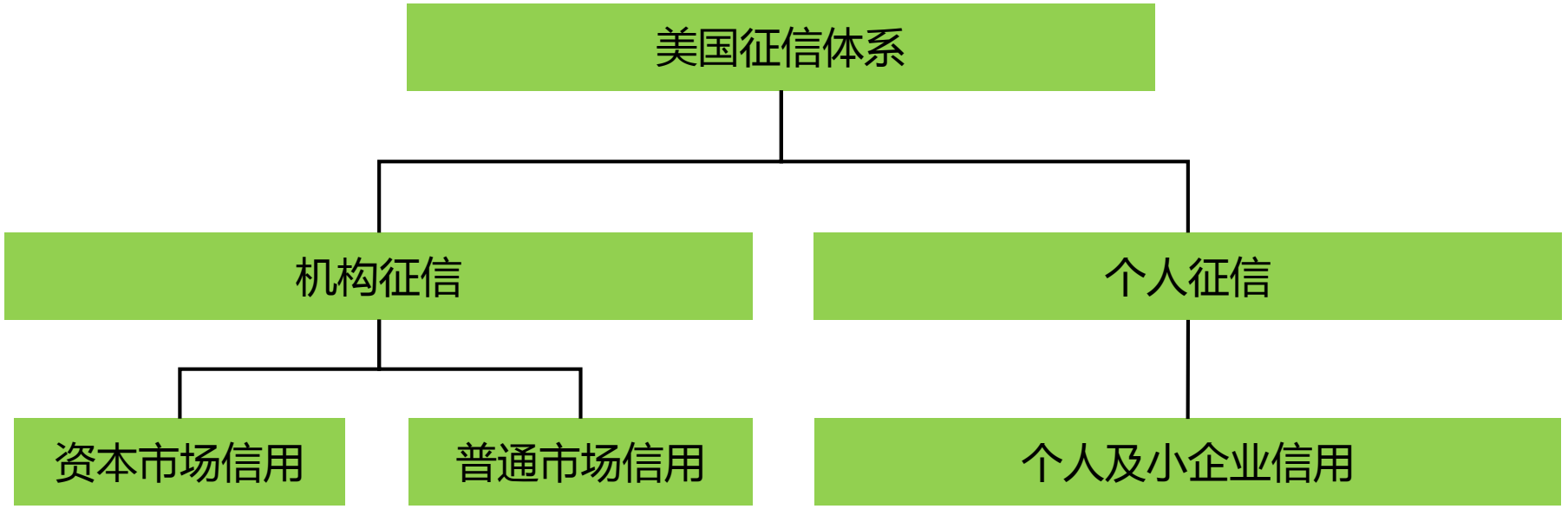
征信解决方案的应用

- 银行
- P2P贷款机构

数据整合和挖掘

- 多维度的数据源整合
- 征信模型建设、服务提供
- 征信解决方案提供

美国著名征信公司



1980年 → 2010年

信息化 市场化

寡头市场 格局稳定

- Standard & Poor
- Moody's
- Fitch

Dun & Bradstreet

- Experian
- Equifax
- Trans Union

Fico VS ZestFinance

- 85%有信贷记录人群
- 银行数据
- 逻辑回归
- 15-50个变量

- 15%无信贷记录人群
- 第三方、网络行为数据
- 机器学习
- 几千上万个变量

国内征信发展历程

- 1988年，第一家信用评级公司“上海远东资信”成立

起步阶段

- **规模小**
- **业务少**

- 1991年，深圳贷款证制度正式颁布执行
- 1991年、1995年，央行总行分两次向全国组织推行贷款证
- 1994年，邓白氏进入中国

平台化尝试

- **央行推行贷款证**
- **各地建设征信体系**

- 2006年3月，央行设立征信中心
- 2013年3月，首部征信行业法规《征信管理条例》开始实施，12月央行制定的《征信机构管理》正式施行

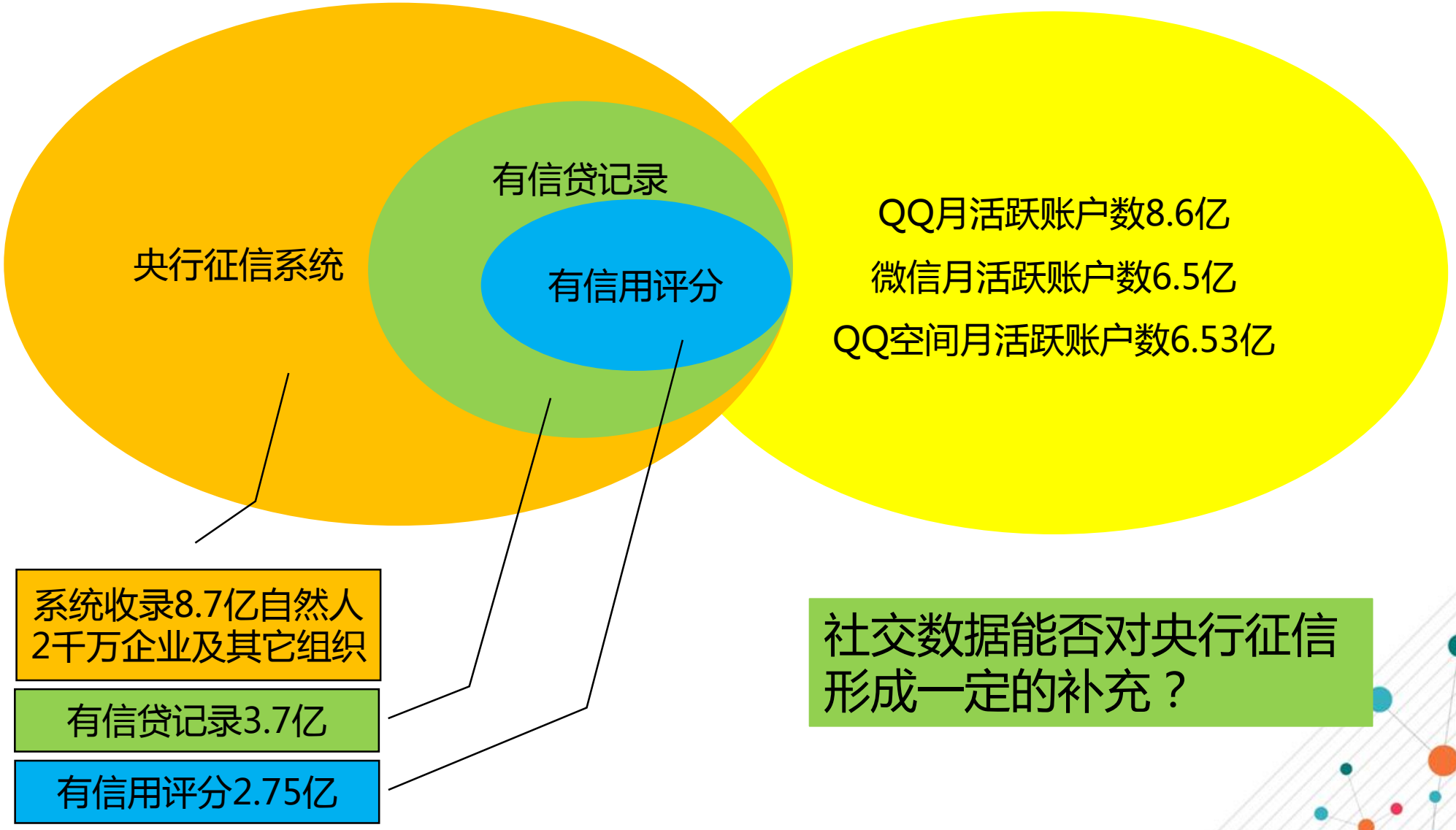
平台化形成

- **央行主导**
- **系统建立**
- **条例规范建设**
- **征信民企成立**

- 央行在2015年1月5日印发了《关于做好个人征信业务准备工作的通知》，要求八家机构做好个人征信业务的准备工作

市场化阶段

- **互联网金融爆发**
- 78家企业征信机构备案完成
- 8家机构获批开展个人征信准备工作
  - **芝麻信用**
  - **腾讯征信**
  - 拉卡拉信用
  - 深圳前海征信中心
  - 中诚信
  - 中智诚征信
  - 北京华道征信



社交数据能否对央行征信形成一定的补充？

社交数据能不能用来做征信？

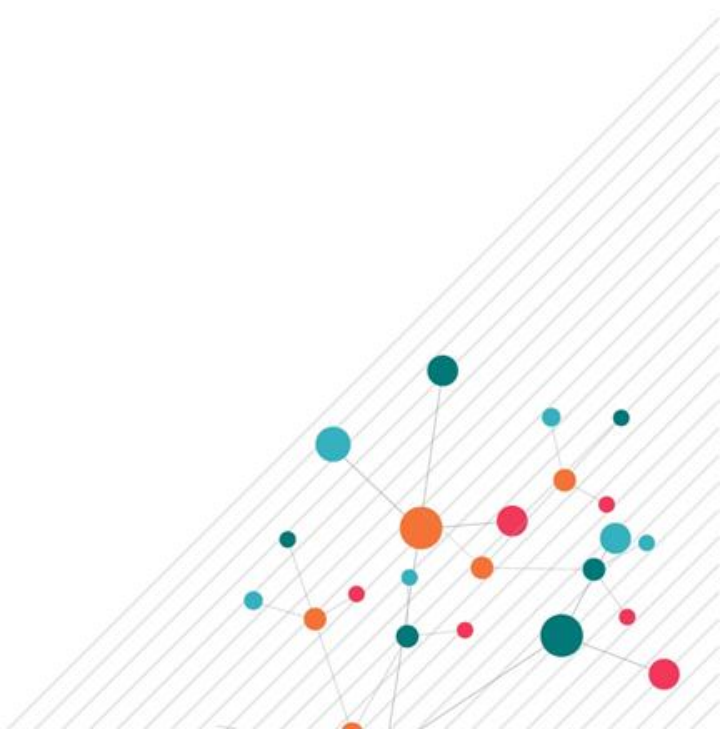
VS

社交数据与信用评级有关系吗？

交易数据天然具备金融属性，社交数据有吗？

社交数据非结构化程度高，怎么挖掘并有效使用？

腾讯社交网络数据



传统征信分析维度

QQ社交数据现状

- **QQ**
  - 月活跃**8.6亿**
  - 最高同时在线**2.39亿**
- **QQ空间**
  - 月活跃**6.53亿**

社交征信SWOT分析

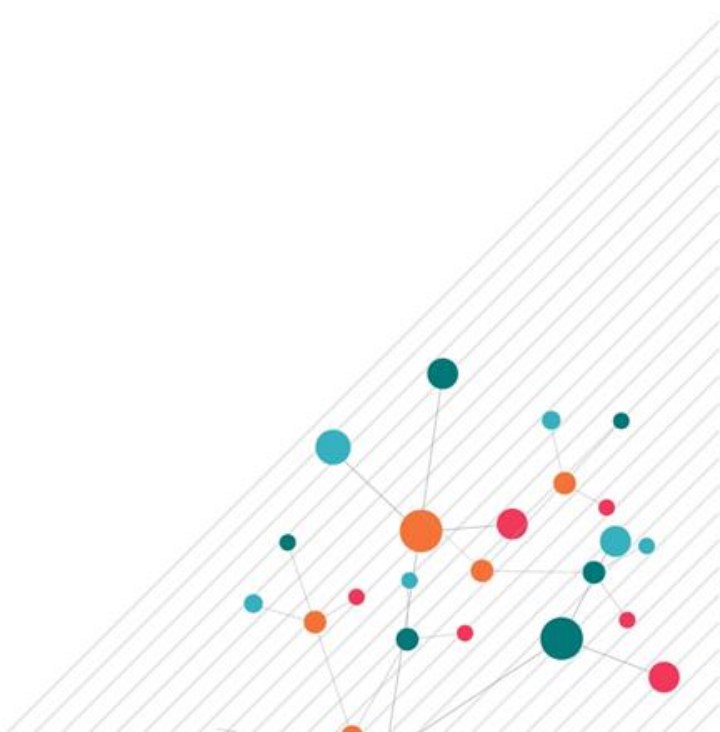
- 覆盖人群广，数据维度多
- 大规模实时数据处理能力
- 算法模型积累较多
- 社交网络数据难以模拟，在反欺诈方面有天然优势

- 互联网金融爆发式增长，征信需求越来越大
- 大量的用户没有被央行征信系统覆盖，空间大

- 缺乏信贷相关核心数据
- 数据类型多，处理难度大
- 缺乏统一的数据标准和规范
- 数据孤岛现象严重
- 账号复杂，一人多号情况多

- 行业对社交征信的认可
- 互联网业务变化快，数据稳定性挑战较大

个体用户画像研究



用户画像主要挑战

1. 如何充分利用腾讯各种丰富的数据资源及之间的联系


社交网络


LBS日志


用户群组


多媒体数据


UGC文本


登录IP

2. 如何使用用户画像适应各种不同的应用场景

广告定向 

推荐系统 

市场营销 

信用评分 

3. 如何高效的处理海量的用户数据（超过10亿的QQ用户，超过千亿级别的各类日志数据）

用户画像解决方案

1. 针对不同的底层数据类型设计特定的挖掘算法，挖掘用户的行为特征，形成底层标签。综合考虑不同数据来源的，形成更上层的抽象用户标签

2. 建立完善的用户画像标签体系结构，从不同维度、粒度对用户进行描述。

3. 搭建用户画像挖掘系统，基于大规模存储和机器学习计算平台，定期对全量用户数据进行计算和挖掘，并提供用户标签的使用和查询服务。

用户画像系统架构

标签应用层

- TDW 离线查询
- HBase 实时查询（理论峰值40w/s）

标签汇总层

- 不同算法、数据来源得到标签进行汇总

模型训练与预测层

- 无监督模型：word2vec, LDA，社区发现
- 半监督模型：标签传播
- 监督模型：LR, Kernel SVM, Random Forest

基于Hadoop，Spark和GraphLab等计算平台

数据处理层

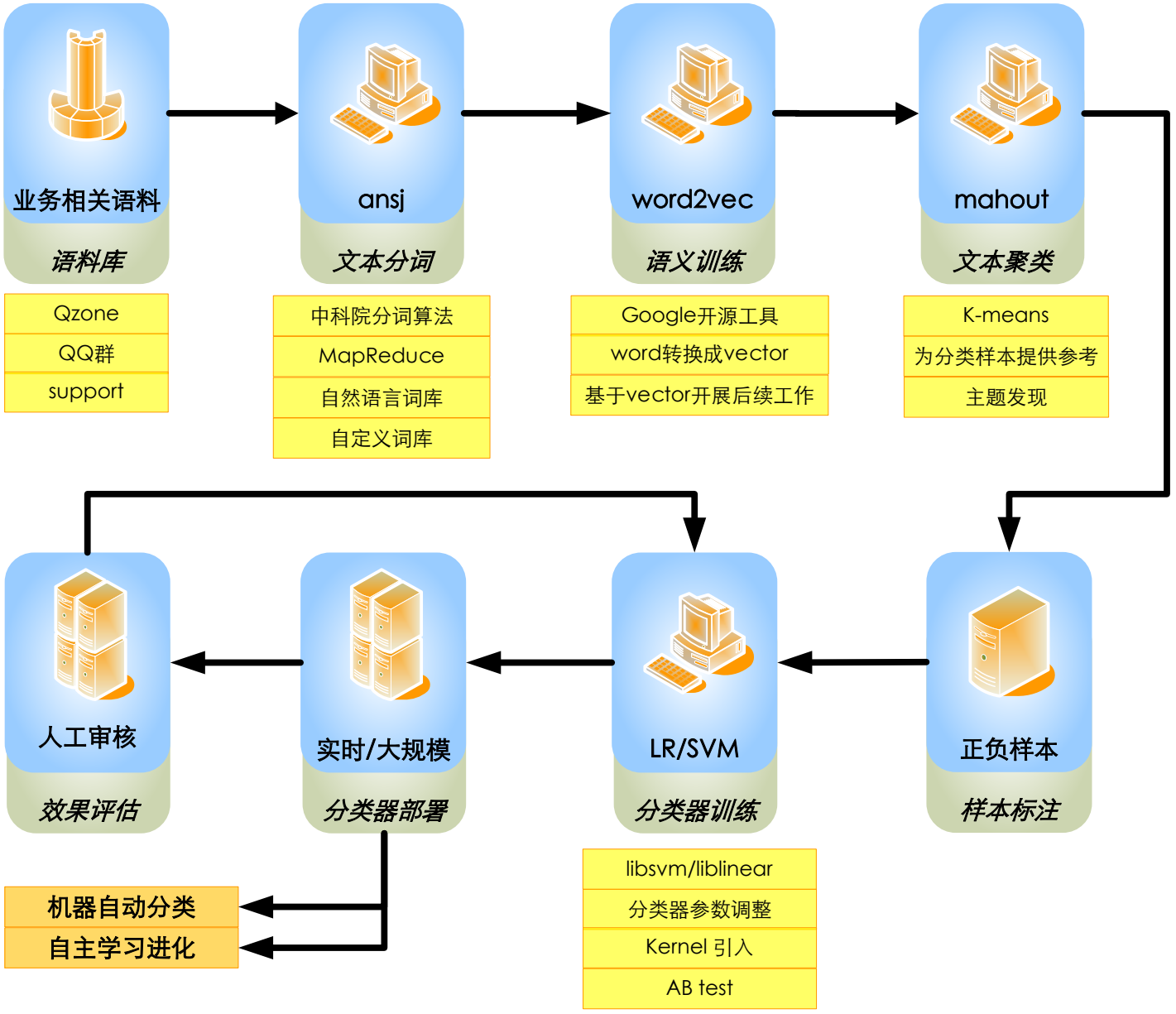
- 结构化数据统计
- 文本分词
- LBS与POI匹配

原始数据层

- 相册说说
- APP文本
- 群文本
- 操作行为
- 关系链
- LBS数据
- 外部数据

TDW数据仓库

用户画像文本挖掘系统



每天处理5600万条说说，3000万相册标题，讲述4500万人的故事

用户画像行业挖掘

思路及问题

思路1：根据用户加入的QQ群文本及其他UGC进行文本分类 → 存在问题：加入群只能反专业业相关兴趣，与职业并无绝对关系

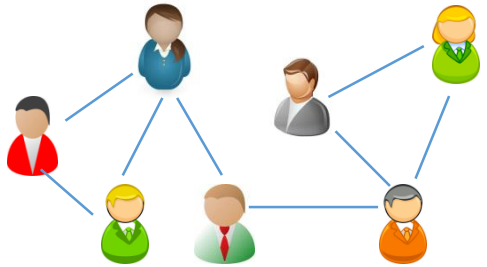
思路2：判断用户工作地点，并根据工作地点推测用户行业 → 存在问题：同一工作地点可能存在多种不同工作行业

思路3：利用同事间好友关系网络进行行业标签传播 → 存在问题：好友关系类型比较复杂，无法确定是否为同事

解决方案

工作地点

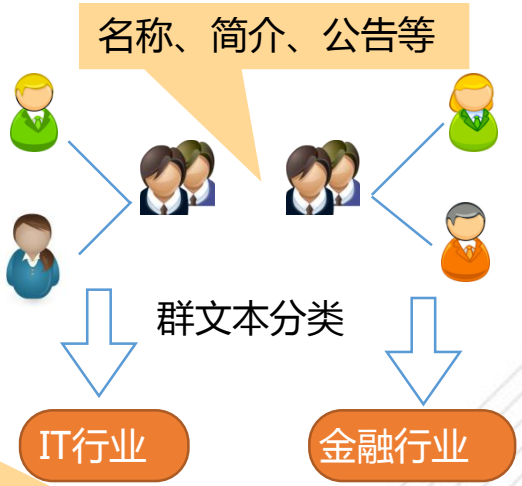
LBS数据挖掘

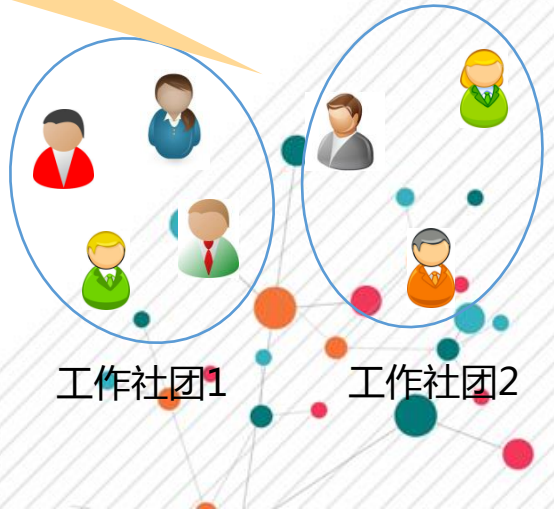


该地点工作的用户及社交网络

Community Detection
(FastGreedy算法)

根据工作社团的特殊性，将部分用户的行业标签扩散给全体社团成员

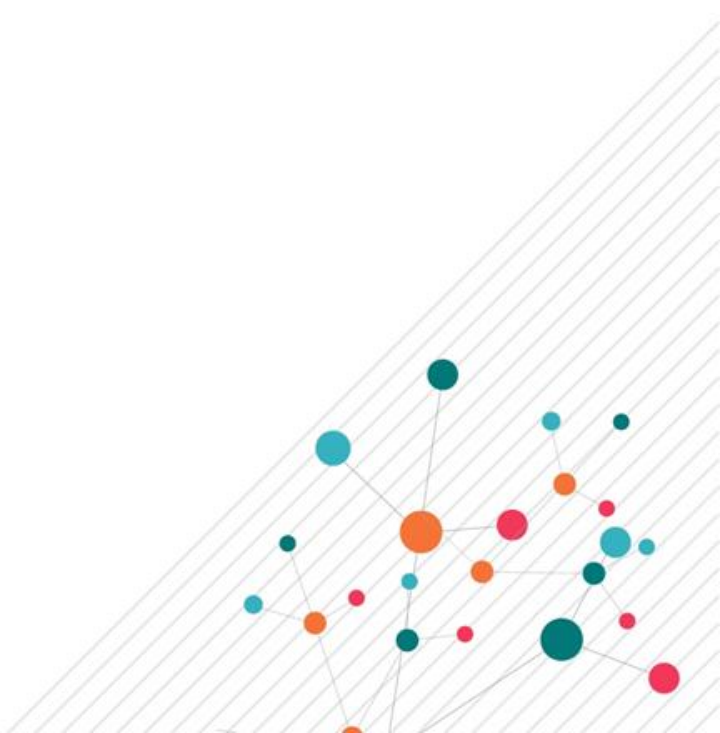




用户画像挖掘结果

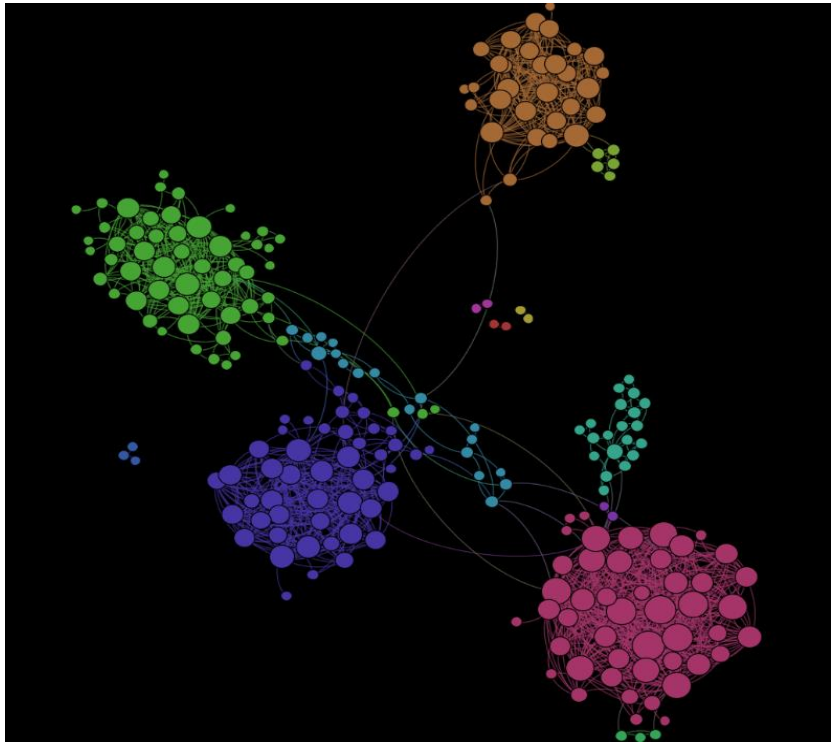
- 结构化数据统计
- 文本分类
- LBS数据挖掘
- 社交网络传播扩散

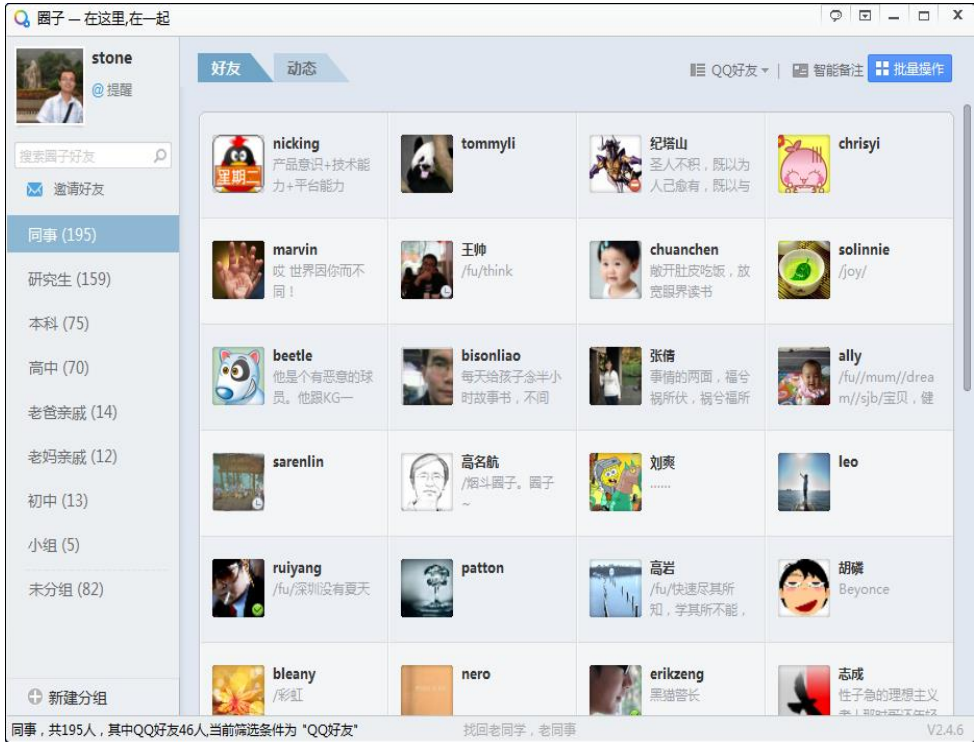
社团圈子研究



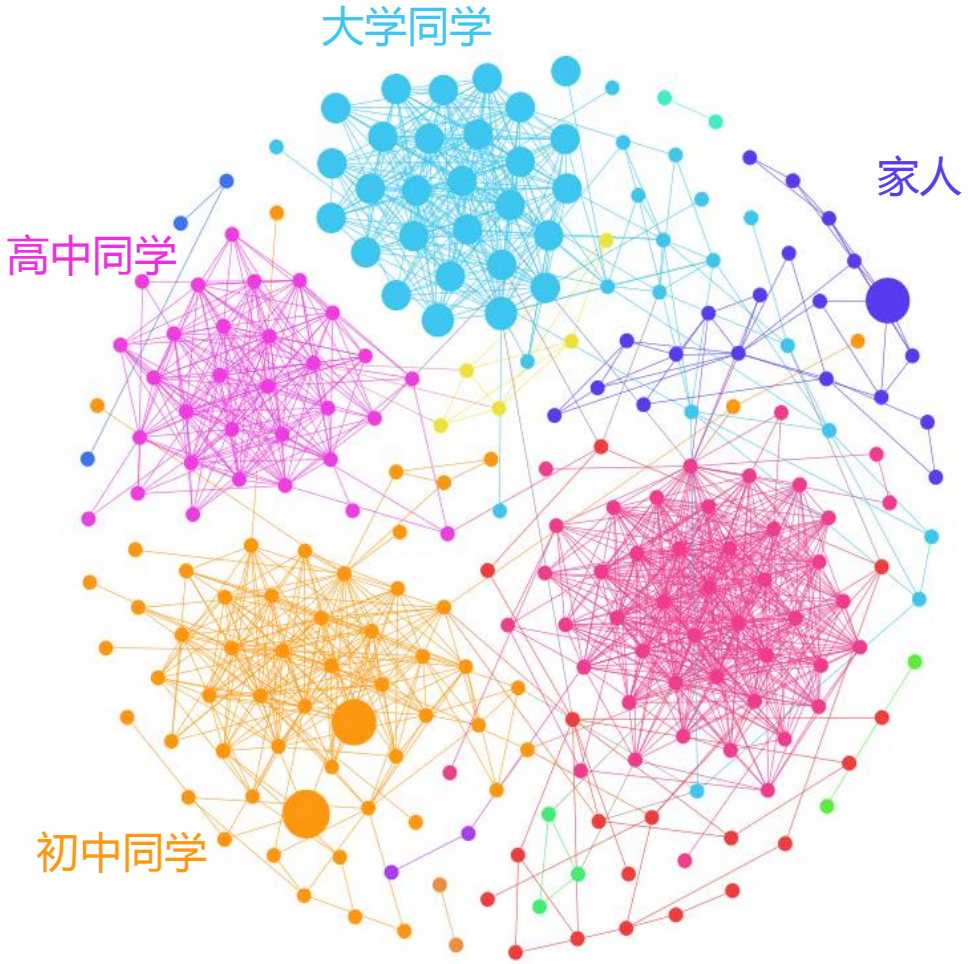
基于个人中心网络的关系链挖掘，QQ圈子

- 首次使用网络拓扑进行智能分圈
- 好友分圈+圈子识别+圈友识别 -> 圈子



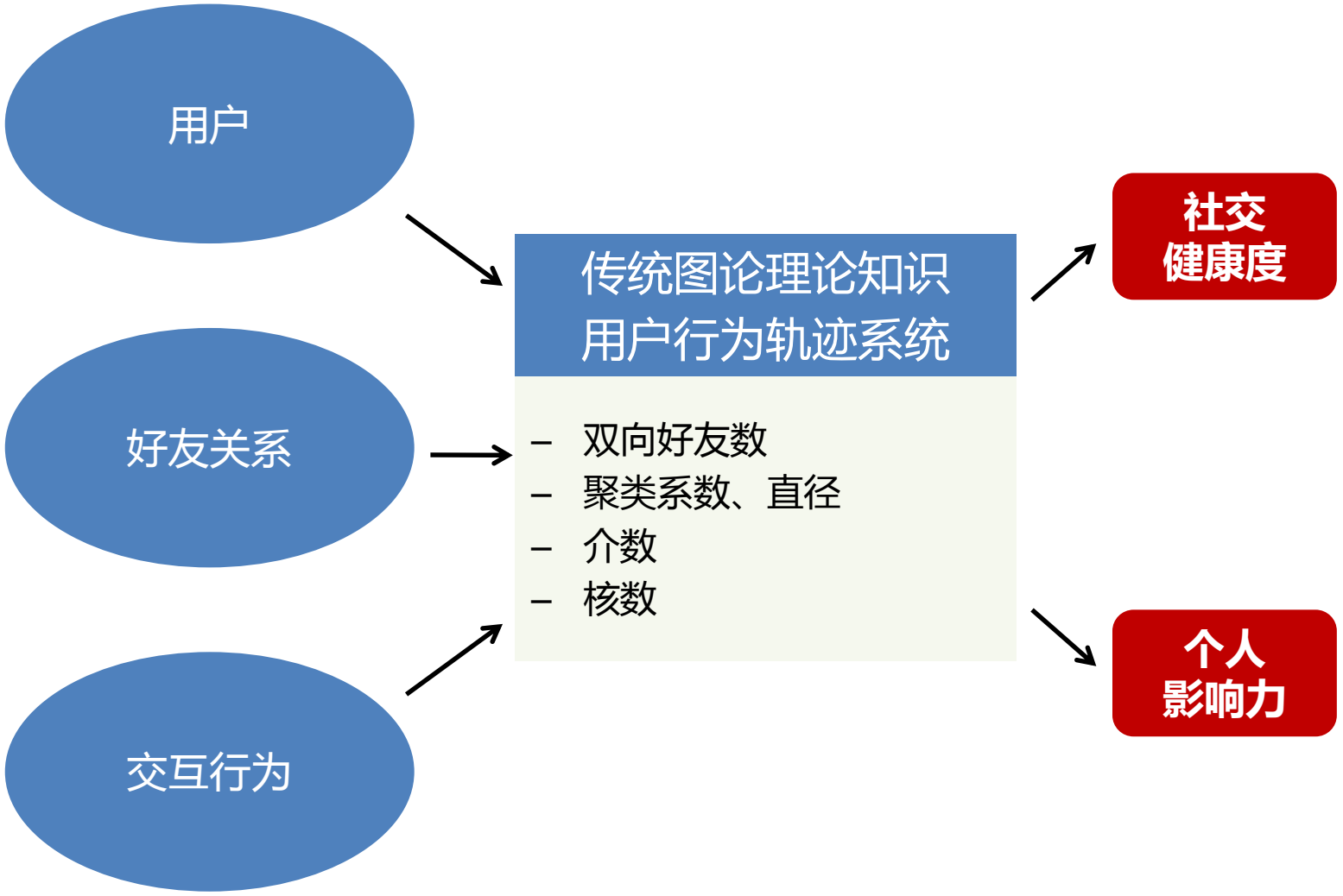


QQ圈子的应用

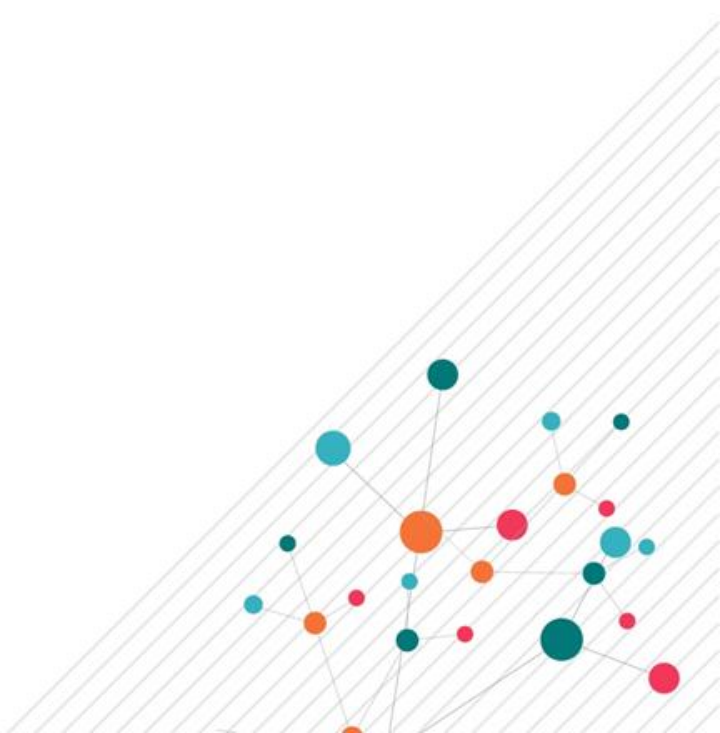


- 基于用户QQ圈子的好友备注分组，挖掘学历信息
- 经验证，QQ学历数据覆盖率为74%，其中普通全日制和脱产的用户准确度均超过90%

社交网络拓扑的应用


用户
好友关系
交互行为
传统图论理论知识
用户行为轨迹系统
– 双向好友数
– 聚类系数、直径
– 介数
– 核数
社交
健康度
个人
影响力

模型建设及应用



自然人计算

- 自然人计算模型的基本假设：

同时在线假设

- 同一设备登录的号码对可能属于同一自然人

互斥假设

- 包括异地互斥和消息互斥，排除异地同时在线、互发消息的号码对

- 自然人计算效果（准确率：84.9%；覆盖率：75.0%）

变量衍生与模型结果

基础变量：

共构造约1200个基础变量

衍生变量：

累计共衍生变量约1万

变量探索：

通过探索及IV分析，确定待选变量约1500个

中间模型：

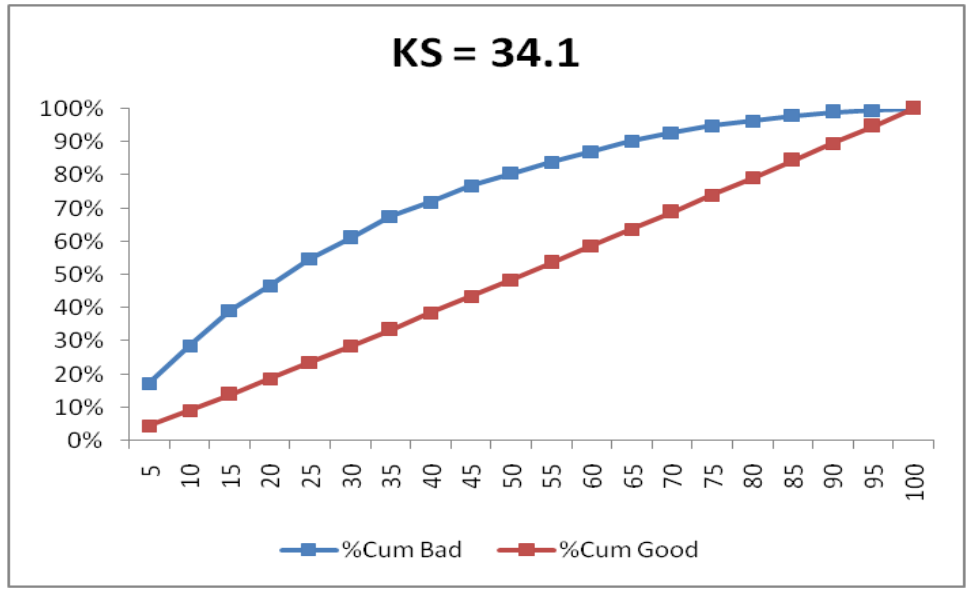
共开发了9个中间模型，模型使用的变量共52个

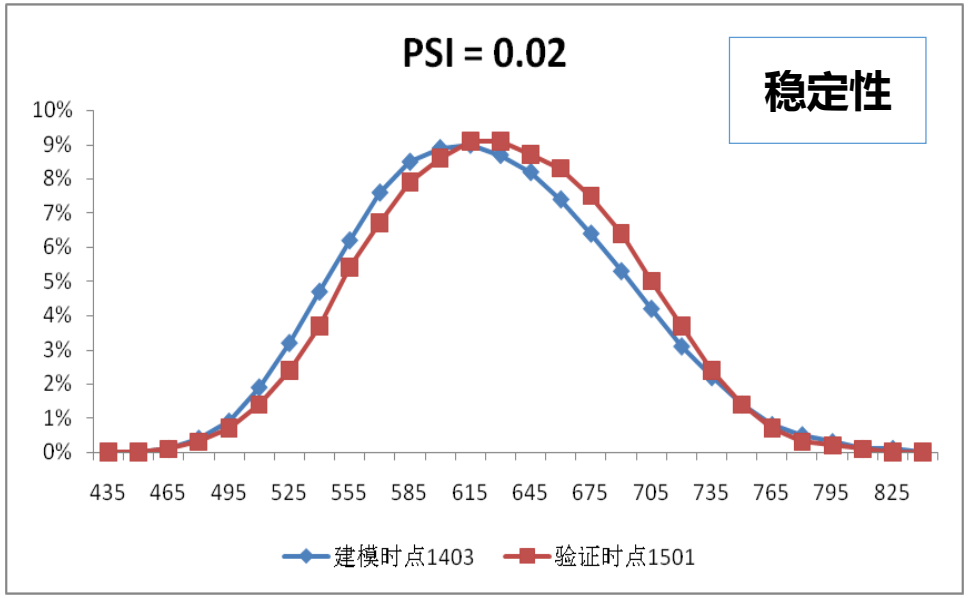
- 关系链组建偏好
- 沟通对象
- 沟通稳定性
- 游戏偏好
- 关系链拓扑结构
- QQ空间
- 群
- 设备安全
- 账户综合信息

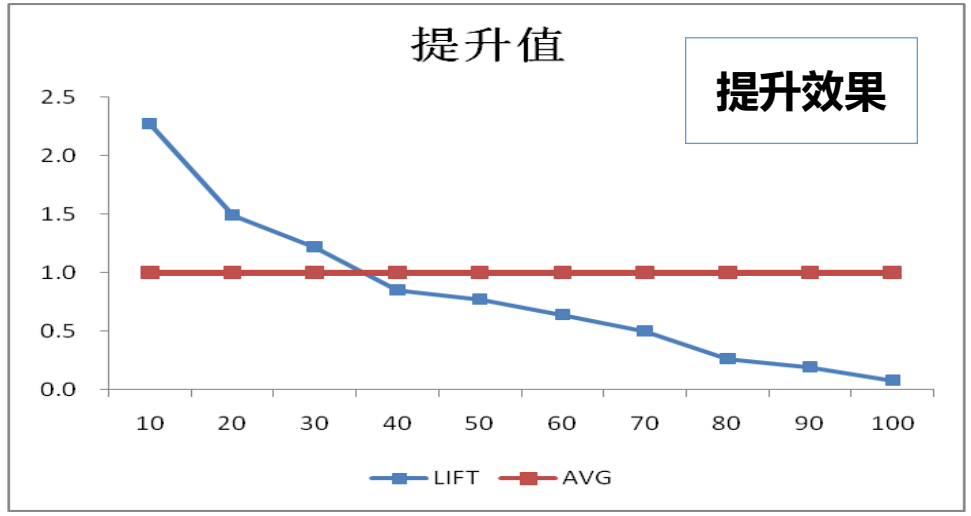
最终模型：

通过LR模型构建最终评级模型

模型整体效果

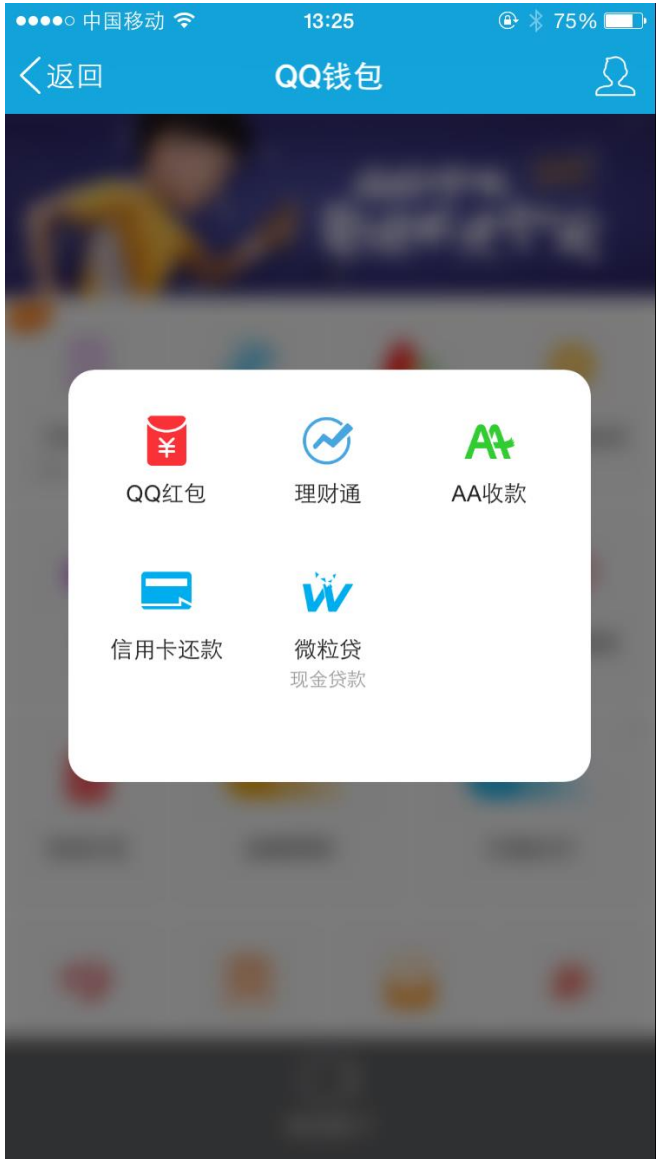






- 模型具有较好的区分能力
- 模型在不同时点分数分布较稳定

微粒贷应用


●●●●○ 中国移动 13:25 75%
返回
QQ钱包
QQ红包
理财通
AA收款
信用卡还款
微粒贷
现金贷款


●●●●○ 中国移动 13:25 75%
返回
微粒贷
借还记录
¥100000
¥50000
¥150000
可借额度
¥199500.00
总额度
¥200000.00
借款
6月21日应还金额
¥28.10
还款详情
常见问题 | 0755-89468888
本服务由微众银行提供 [珍惜信用 按时还款]


●●●●○ 中国移动 13:28 74%
微信
招商银行
下载
详情
13:27
交易提醒
5月21日
尊敬的客户:
您尾号9878的招行一卡通入账人民币500.00元
交易时间：05月21日
交易类型：汇款到账
交易金额：人民币500.00元
付款人姓名：微粒贷
付款人账号：0183
☆手机转账0费用☆ 每日高达20万额度，使用手机银行转账汇款，立享免费。点击全文下载
详情
我
发现
无卡取款

Thank you