用户画像–大数据时代下的用户洞察

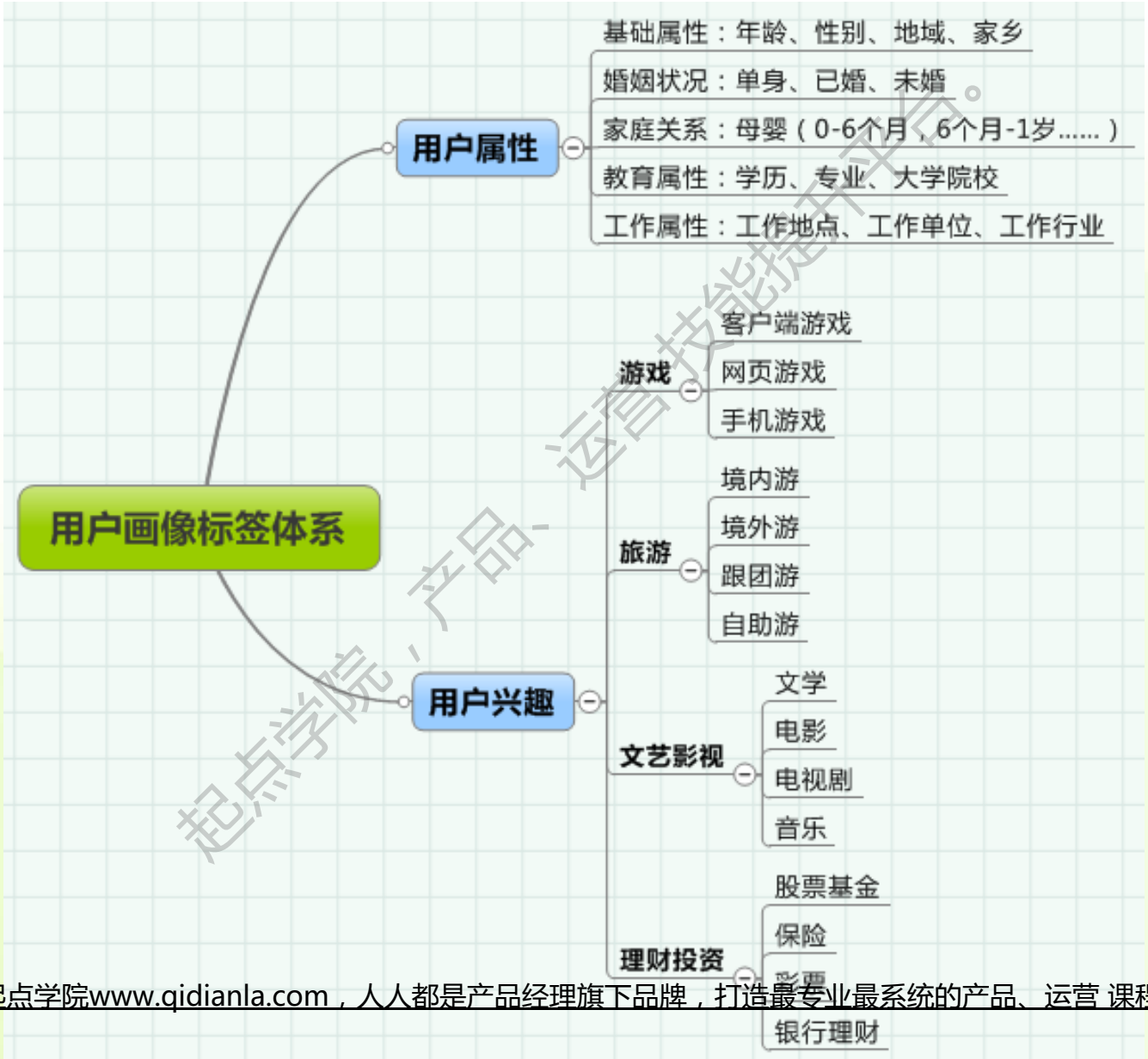
刘黎春
SNG运营部/数据中心
May 2015

目录

SNG数据现状

- QQ
  - 月活跃8.4亿+
  - 最高同时在线2亿+
- QQ空间
  - 月活跃6.5亿+

用户画像体系



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。

用户画像挖掘的基本框架

单一数据源挖掘

数据源

结构数据统计

文本分类

社交网络分析

LBS数据挖掘

… …

社交网络

底层标签

底层标签

标签在社交网络中的扩散

底层标签

底层标签

高层标签

从底层标签挖掘高层标签

文本挖掘系统

QQ群

QQ空间

… …

文本预处理 → 特征提取 → 文本分类

- 中文分词
- token抽取

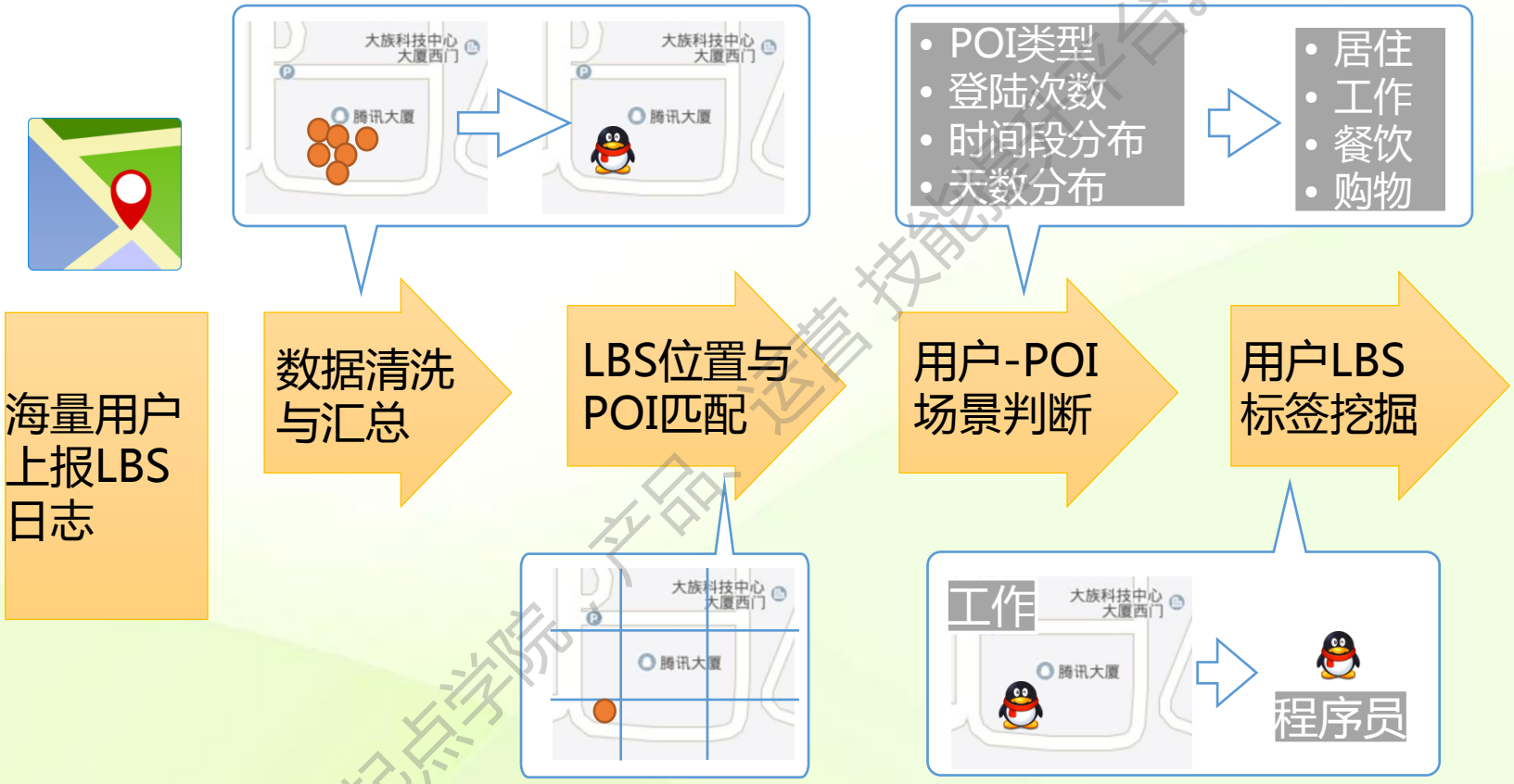
- tf-idf
- LDA
- word2vec

针对短文本特点，利用LDA与word2vec进行语义扩展

- logistic regression
- Kernel SVM
- Neural Networks

利用非线性分类器对神经网络得到的特征向量进行分类

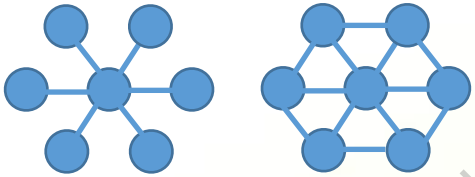
基于LBS数据的用户画像挖掘



社交网络与用户画像

用户在社交网络中的行为反应出现实生活中的某些特质：

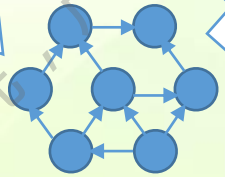
局部聚类系数：
(local clustering coefficient)



反映用户与好友关系的稳定性及QQ用户交友的主要目的

社团影响力
PageRank得分

反映用户在社交网络中人脉的丰富程度或重要性

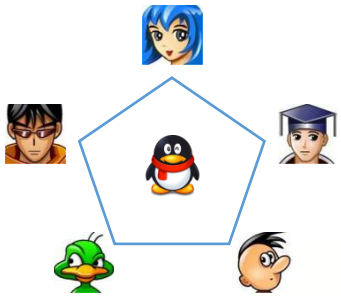


根据用户间的重要程度，将无向图转化为有向有权重的好友关系图

利用Pagerank算法对有向图中的所有节点进行排序，得到不同节点的影响力得分

基于社交网络的标签扩散

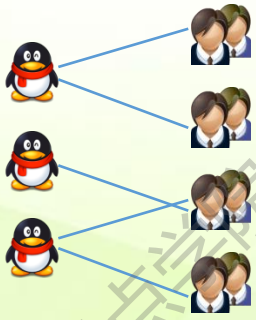
好友关系网络下的标签传播



算法改进：
- 好友关系类型对传播的影响
- 好友关系的稳定性对传播的影响

算法应用：
- 用户基础属性优化，如年龄
- 用户属性扩散，如职业、学校等

群-用户二部图下的标签传播



算法改进：
- 针对QQ群的特殊场景设计标签传播算法，提升传播效率和准确度

算法应用：
- 用户属性扩散，如职业、学校等
- 用户兴趣扩散，如文艺、体育等

不同数据源的融合 – 职业挖掘

如何判断一个用户工作所在的行业

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

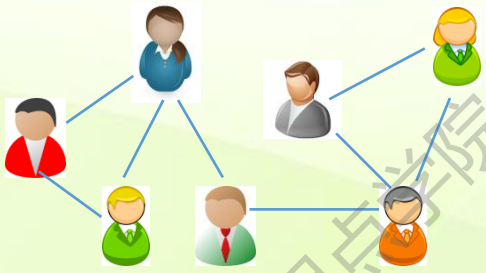
思路3：利用同事间好友关系网络进行行业标签传播

存在问题：好友关系类型比较复杂，无法确定是否为同事

不同数据源的融合 – 职业挖掘

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)

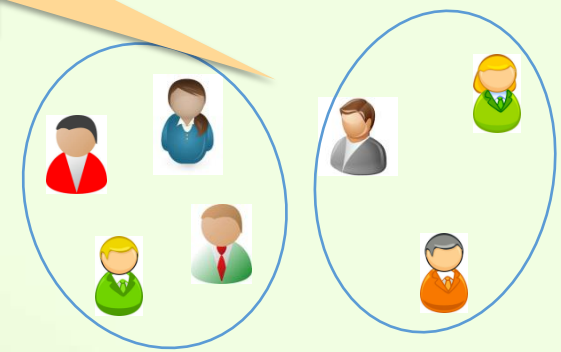
名称、简介、公告等

群文本分类

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

IT行业

金融行业



工作社团1

工作社团2

计算平台与系统部署

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

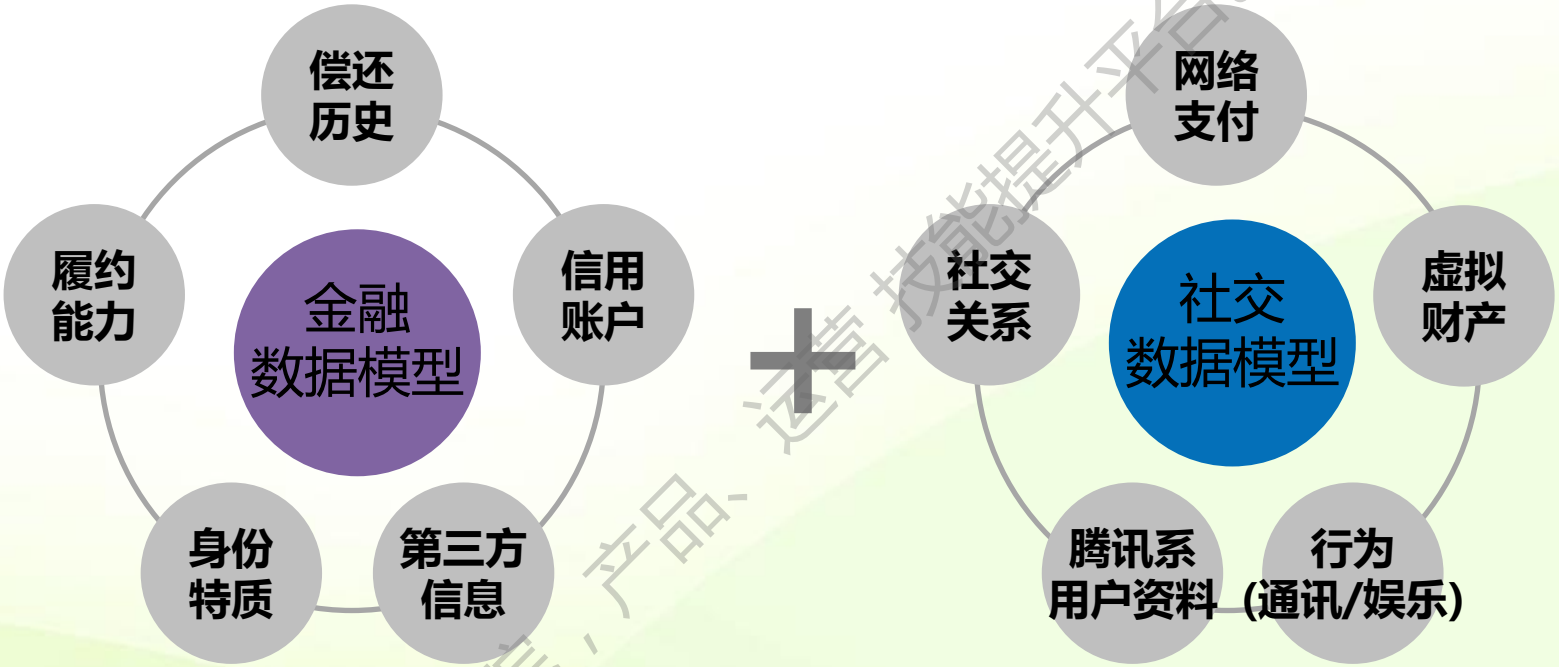
原始数据层

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

用户画像应用－广点通定向投放

用户画像应用－腾讯征信



数据银行 | 机器学习 | 用户画像 | 统计学

谢谢！