China · Beijing

# 基于文本数据的用户画像实践

丁若谷

明略数据技术合伙人

明略数据 MININGLAMP

ArchSummit 全球架构师峰会 2016

[ 北京站 ]

主办方 Geekbang. 极客邦科技 InfoQ

# 大纲

用户画像概述

大数据环境下的技术挑战

基于文本数据的画像系统精益演进

人工智能和自然语言处理技术

明略用户画像实际效果

# 用户画像概述

- 基于什么数据
- 能产生什么价值
- 常见数据和应用场景

# 最简单的用户画像系统

## 场景举例

- 某银行希望判断某客户是否“土豪”，决定是否发放黑卡

## 实现方式

- select balance, education from tbl_acc_info where id= ‘13888888888’；
- 前端对接查询系统

# 当数据变得更大

## 存储

- 分布式存储（HDFS）
- 分布式数据仓库（Hive）

## 计算

- 预先计算好复杂标签
- NoSQL数据库提供查询（HBase）

# 重新考虑场景

## 场景举例

- 某银行希望针对“土豪”客户，批量发放黑卡

## 实现方式

- select id, name from tbl_acc_info where balance > 100000000 and education < 3;
- 前端对接查询系统

# 当数据变得更大

## 存储

- 源数据是如何组织的？

## 计算

- 逐个用户进行计算vs逐个标签进行计算？
- 标签预期如何分布？

# 明略的选择

## 明略服务客户特点

- 全国性客户为主，用户量在十亿量级
- 需求多样，标签数量在2000左右
- 数据较为稀疏

## 技术选择

- 按用户组织数据

# 技术挑战：标签组合反查

## 场景举例

- 余额大于100万
- 教育程度初中以下
- 居住在北京市朝阳区
- 有高端会所消费
- ……

## 需求分级：数量需求，列表需求

# 应对方案

## 基准方案

- 扫描全表：10分钟量级
- 建立查询缓存

## 针对数量需求的优化

- 扫描抽样表：1秒量级
- 预建cube：划定n个标签的范围，组合最多m个标签，预先计算好各种组合的计数

# 应对方案（续）

基准方案

针对数量需求的优化

针对列表需求的优化

- Hive+Parquet+Impala
- Top-k缓存

# 文本数据概述

## 场景举例

- 呼叫中心语音转文本
- 贷款调查报告文本
- 社交媒体文本

## 非结构化数据和结构化数据的差异

- 维度更高
- 需求不明确，研发更困难

# 基于文本数据的画像系统精益演进

## 画像算法

- 关键词-正则-逻辑规则

## 规则管理和探查

- grep+配置文件
- 基于Web的规则管理和探查系统

# 当数据变得更大

## 画像算法实现方式

- Python单机版
- MapReduce
- AC自动机多模匹配

## 规则探查

- 基于抽样数据
- 引入ElasticSearch，基于全量数据

# 新问题：营销内容的区分

## 场景举例

- 转发微博营销内容，影响画像质量

## 应对方案

- 人工标注训练集，建立机器学习模型
- Spark实现

# 锤子在手，天下我有？

## 行为

- 希望将营销内容的区分推广到更多标签
- 人工标注5个标签的训练数据，建立机器学习模型

## 结果

- 标注质量失控
- 数据稀疏性导致样本不均衡
- 引入第三方数据改进效果

# 技术挑战：命名实体识别

## 场景举例

- 客户希望根据贷款调查报告的描述，区分不同行业贷款客户

## 基准方案

- 国民经济行业分类(GB/T 4754-2011)
- 人工标注10万条贷款调查报告，将其归类至国家标准
- 建立机器学习模型

# 应对方案

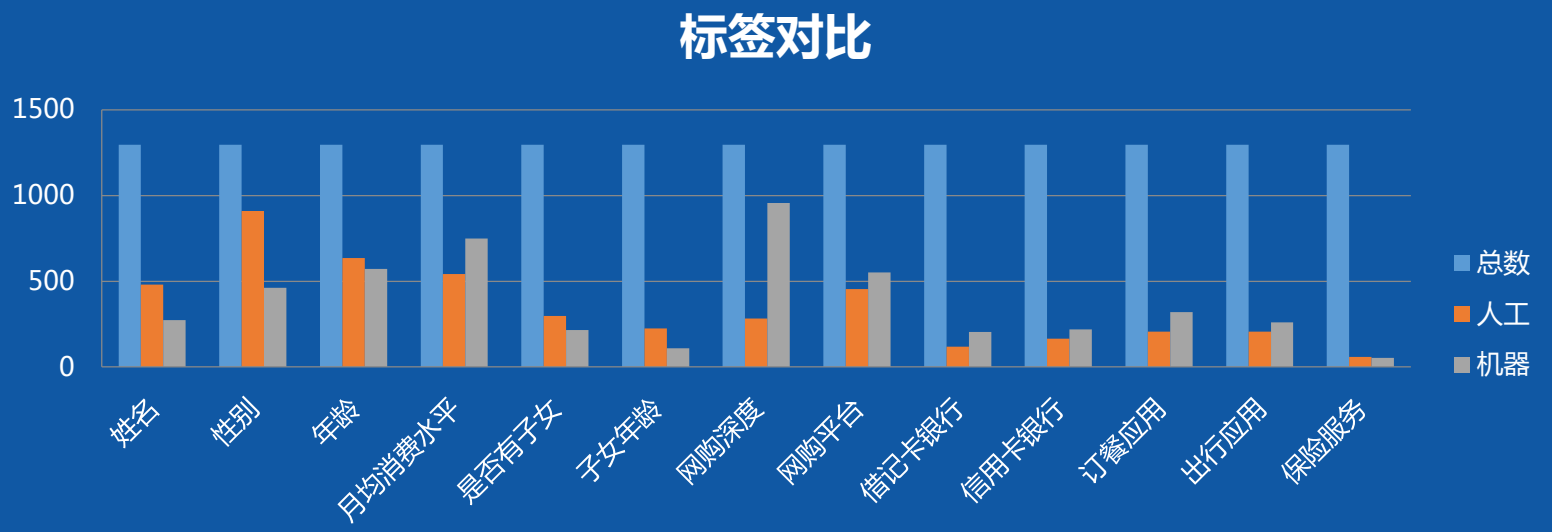
## 基准方案的缺陷

- 人工标注成本过高，无法发现新的实体

## 自然语言处理方法

- 文本分词，统计词频
- 在训练集上采用LSI模型进行文本分类
- 建立PHMM模型，估计参数
- 在测试集上匹配不同模型

# 明略用户画像实际效果


标签对比
1500
1000
500
0
姓名
性别
年龄
月均消费水平
是否有子女
子女年龄
网购深度
网购平台
借记卡银行
信用卡银行
订餐应用
出行应用
保险服务
总数
人工
机器

# THANKS

ArchSummit
全 球 架 构 师 峰 会 2016

[ 北京站 ]

主办方 Geekbang. 极客邦科技 InfoQ