TiD2016
质量竞争力大会
软件研发顶级盛会

下一代软件研发
SOFTWARE DEVELOPMENT
更高 更新 更深

下一代软件工程中的数据智能研究

DJ Ning

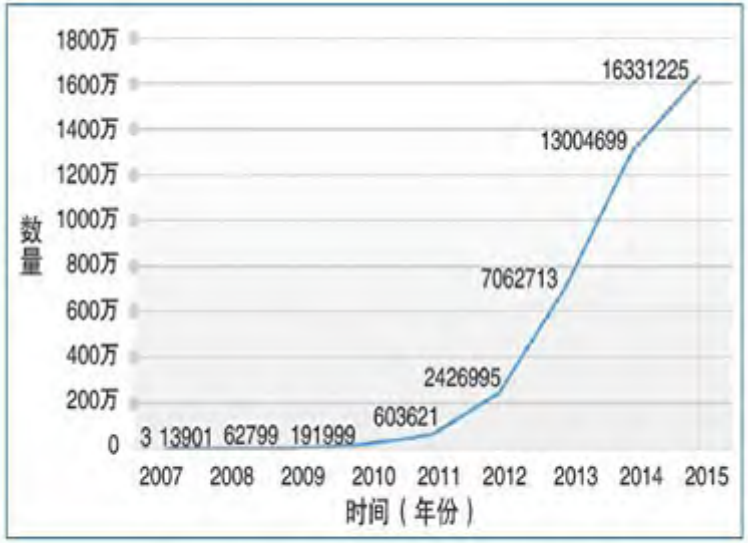
教授级高工

中科院上海高等研究院

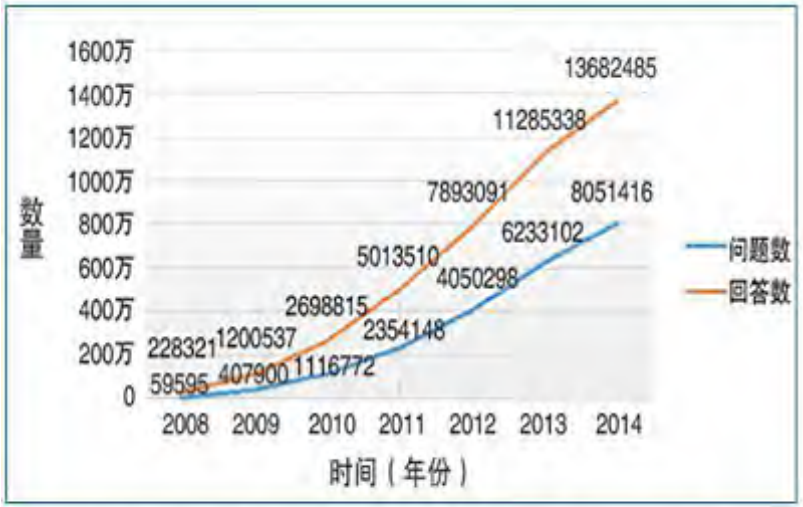
汇报提纲

- **开源软件开发**
- 以群体智能和数据智能为核心的CDDevOps
- 软件全生命周期的数据智能

开源社区的发展



GitHub的代码库/项目增长图



StackOverflow社区的问答数量增长图

项目研究社区&统计数据背景

国外现状

- 开源软件的推崇推动着GitHub、SourceForge、StackOverFlow等开源社区的兴起
- 各种团队协同开发工具使得开发过程被记录
- 软件产品从单一的关注开发过程，到全面关注开发、运营和用户使用过程的协同创新

国内现状

- 华为等大型企业越来越重视内部开源的开发模式
- 学术界逐渐关注在软件工程领域的数据挖掘与分析
- 开发、运营过程数据监控工具的出现

端网云基础设施和大数据的发展，使得人们越发重视软件生命周期数据收集，这正改变着软件工程过程的可见性、智能程度和软件创新的方法

头脑风暴：开源软件成功的关键要素

- 社会化(social)程序员主动参与在线开发社区
- 良好地参与和自组织文化
- 技术和众包内容的快速创造和扩散
- 软件知识正在成为公共资源
- …

汇报提纲

- 开源软件开发
- **以群体智能和数据智能为核心的CDDevOps**
- 软件全生命周期的数据智能

分组讨论：下一代软件工程方法

TiD2016

需求	架构	开发和测试	部署和分销	运营
涌现式需求；基于场景的需求分析；用户参与产品定义和设计	端网云架构模式；关注宏观架构策略；关注战略层面重用	支持云+端的多模软件开发、测试环境；社会化协同模式；	应用商店+服务消费方式 免费+分享模式	云+端的运营模式；社区生态环境的运营；关注U、R、E&E

下一代软件工程的主要变化和趋势

下一代软件工程的主要变化：				
	互联网正在成为一个开放的开发环境	互联网成为一个架构设计和执行的基础设施	互联网正在成为软件分发和销售的基础设施	软件用户处于随时可能移动且随时在线的状态

软件开发主要趋势：

1. 从严格的遵从软件开发过程到面向目标的智能汇聚
2. 软件定义、设计、开发、运营和业务价值交付之间的界限逐渐退化
3. 软件产品的成功交付建立在团队动态互动和协作的群体智能基础上
4. 基于开放社区的软件创新生态环境的形成
5. 能耗、安全和用户体验等非功能性需求日益重要

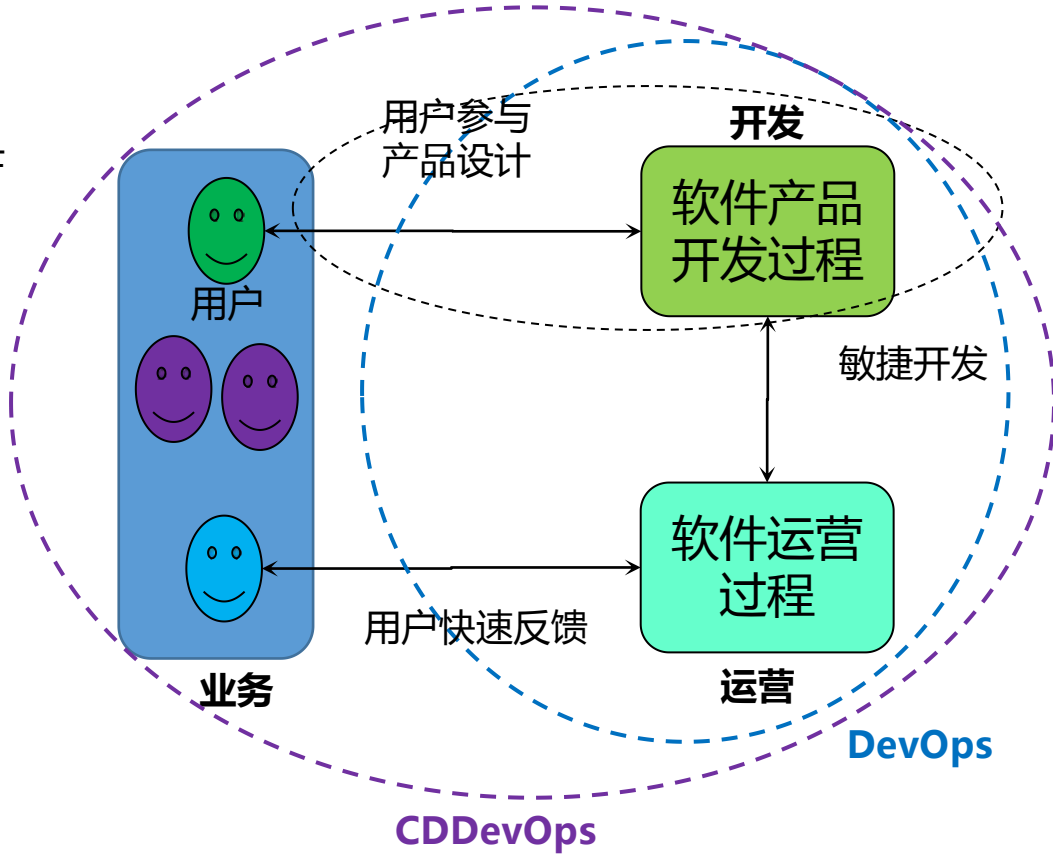
探究下一代软件工程方法： CDDevOps

敏捷开发：更加关注个体与交互、可用的软件、用户参与和快速响应变化

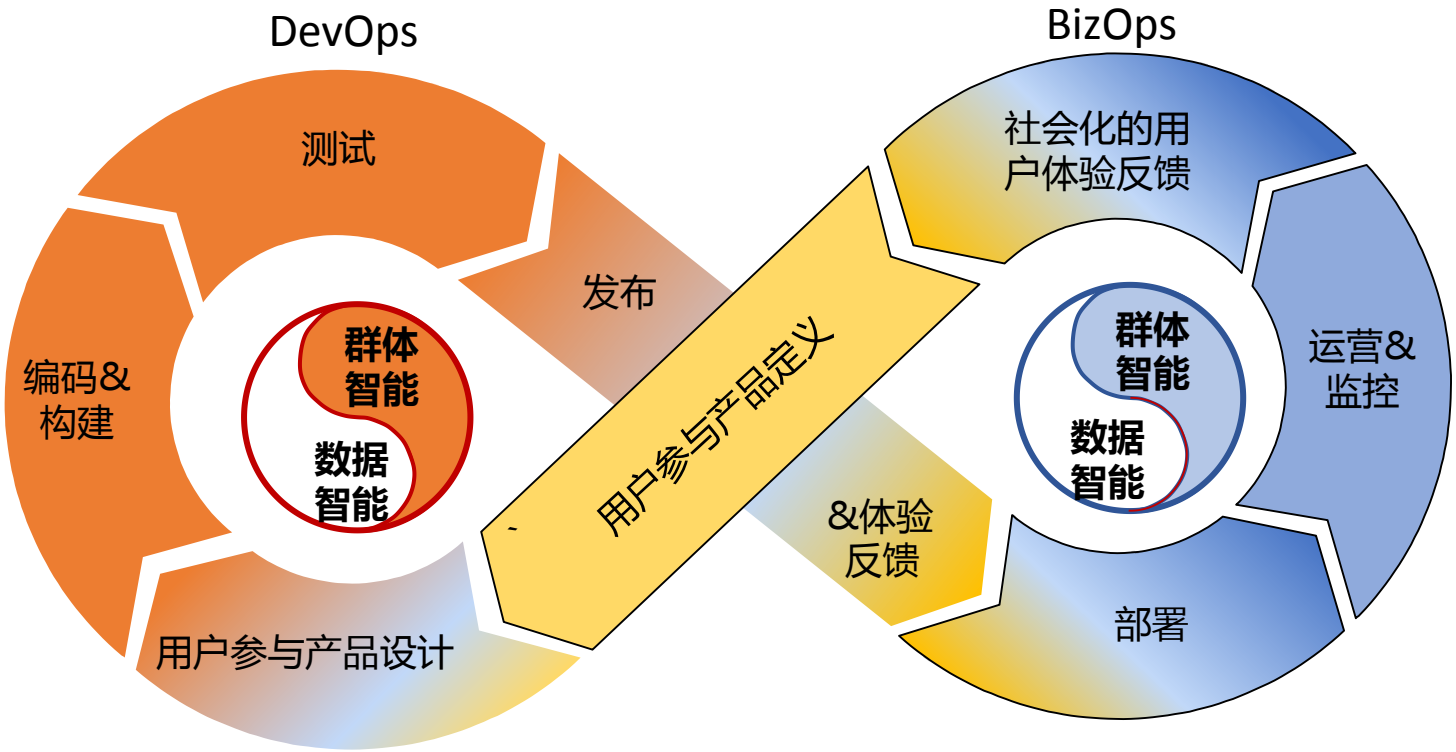
DevOps: 通过打通开发运营实现软件产品的快速发布和反馈。

基于**群体智能**和**数据智能**的DevOps方法（简称CDDevOps）：

- C – CI - Collection Intelligence
- D – DI – Data Intelligence



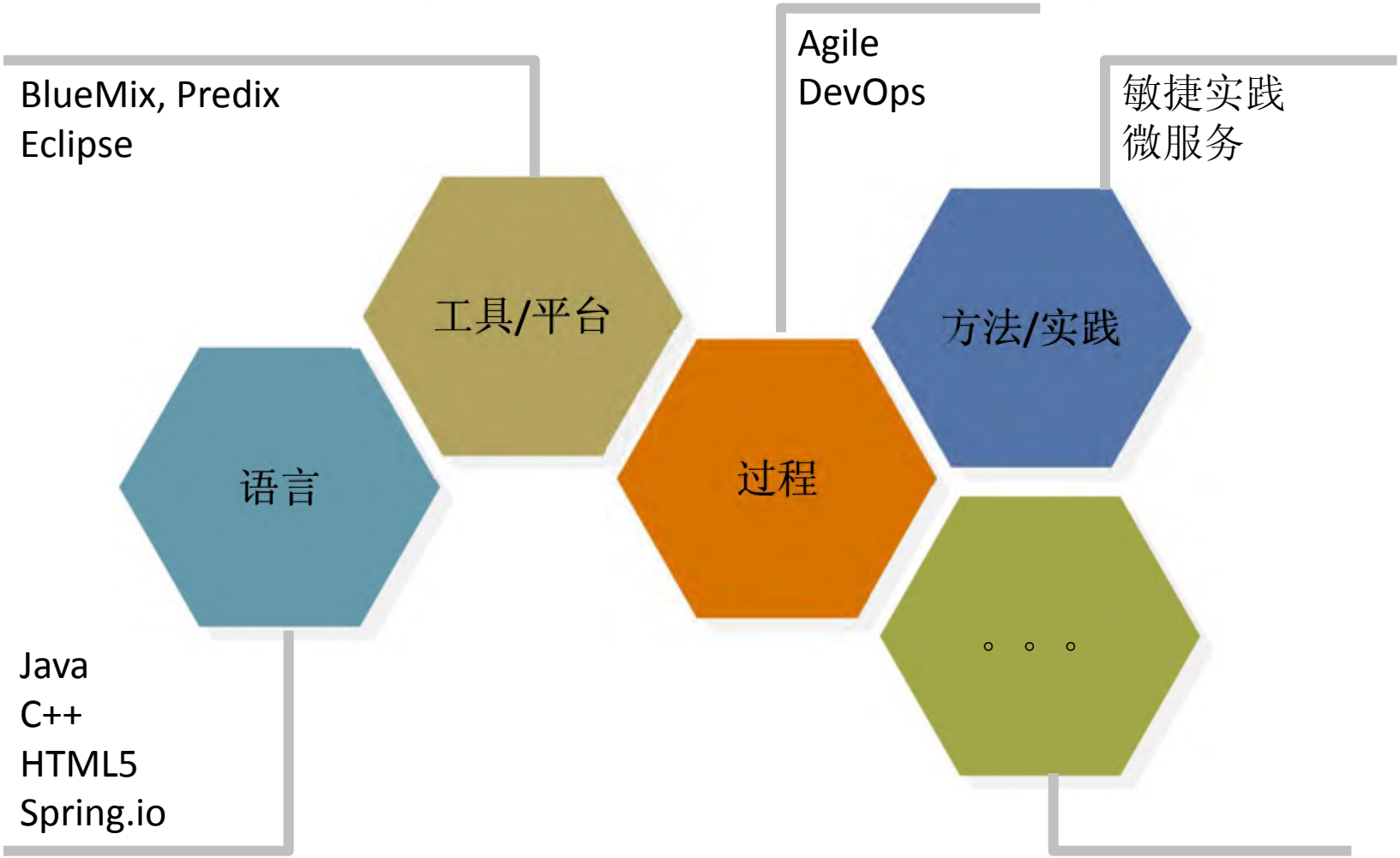
CDDevOps开发过程模型



两个主要创新点：

1. **CDDevOps模型**中的**群体用户协作智能**
2. **CDDevOps模型**中的基于**数据分析基础上的数据智能**

头脑风暴：探究下一代软件工程



汇报提纲

- 开源软件开发
- 以群体智能和数据智能为核心的CDDevOps
- **软件全生命周期的数据智能**

CIDevOps开发过程的感知

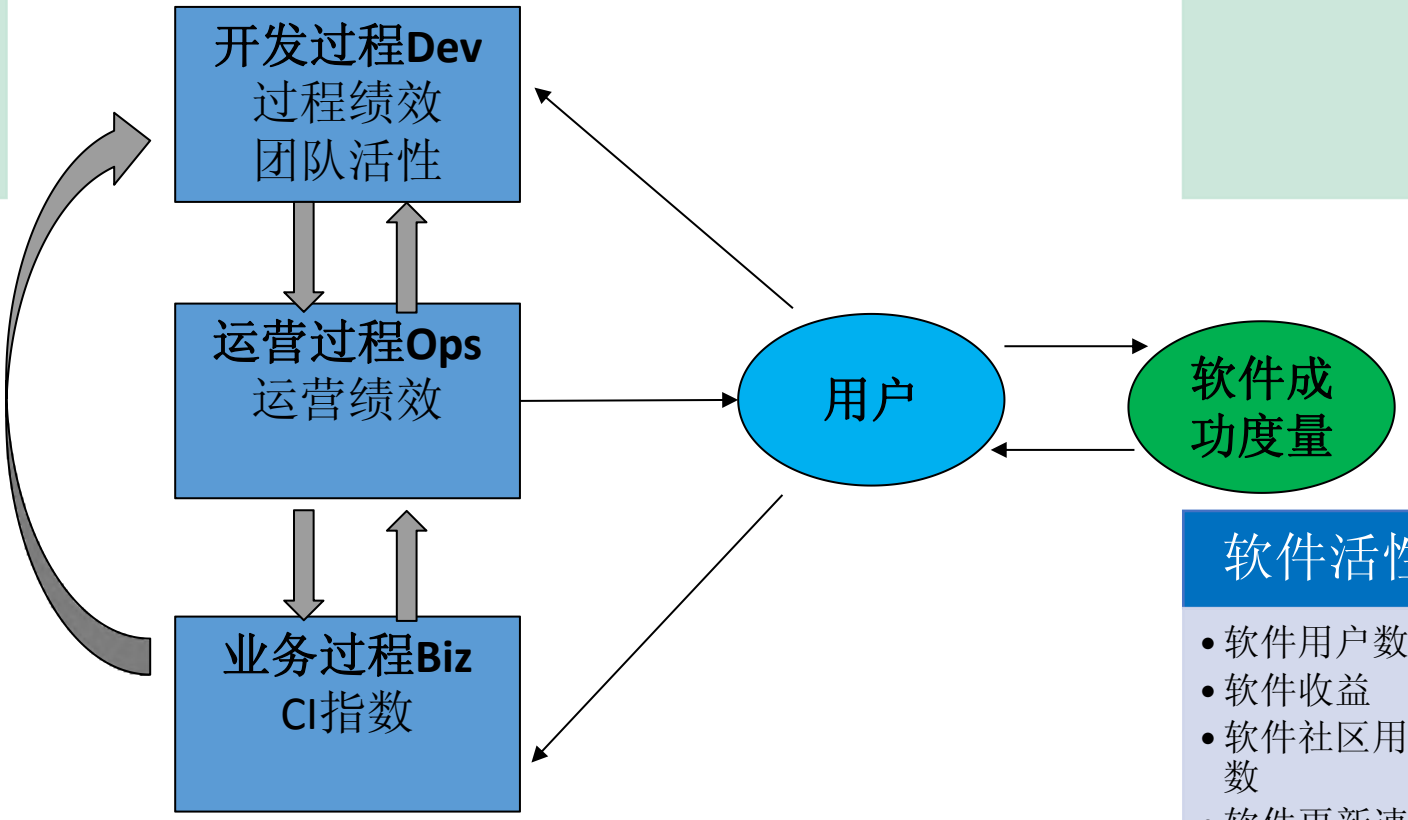
TiD2016

过程绩效

- 代码提交量(行数)
- 文档提交量
- 缺陷修复速率
- ……

运营绩效

- 用户增长速度
- 收益增长速度



团队活性

- 代码提交数
- 代码更新率
- 开发者社区协作紧密程度
- 团队自组织程度
- ……

软件活性

- 软件用户数
- 软件收益
- 软件社区用户数
- 软件更新速度
- ……

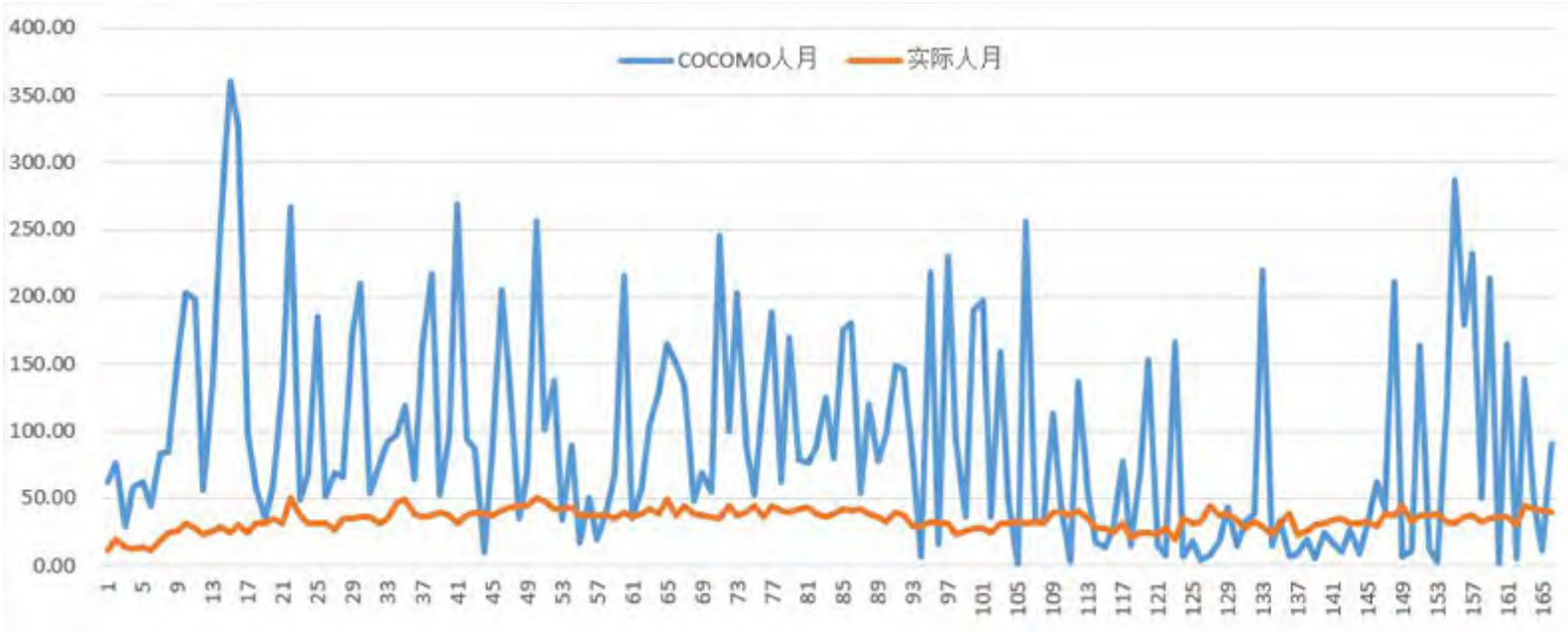
研究内容

团队活性	过程绩效	运营绩效	软件活性
• 代码提交数 • 代码更新率 • 开发者社区协作紧密程度 • 团队自组织程度 • ……	• 代码提交量(行数) • 文档提交量 • 缺陷修复速率 • ……	• KOL挖掘 • 用户特征挖分析 • 传播路径分析 • ……	• 软件用户数 • 软件收益 • 软件社区用户数 • 软件更新速度 • ……

分组讨论：数据智能在软工中的应用

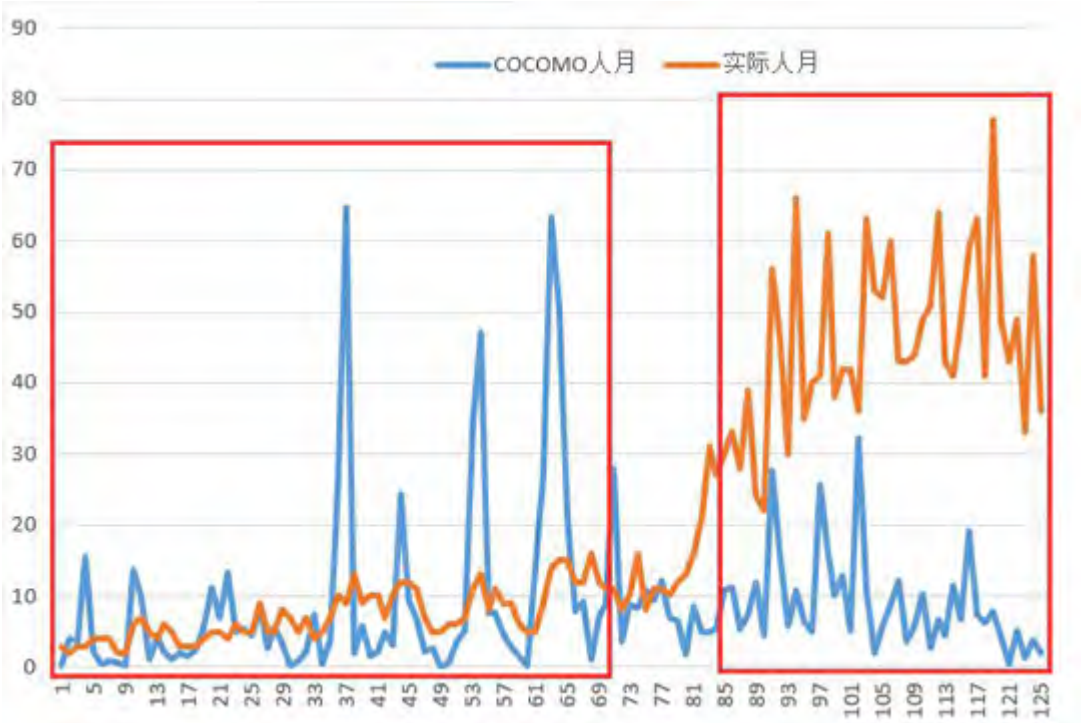
- 客户理解 – 用户画像、市场分析
- 需求获取 – 用户行为分析、文本分析

开源社区软件开发团队绩效数据分析实践



Mono项目COCOMO人月与实际人月

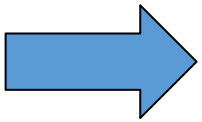
开发团队绩效数据分析实践



Django项目COCOMO人月与实际人月

开发团队绩效数据分析实践

	项目数（个）
研究项目总数	53
显著相关	32
相关性大于0.5	10
相关性大于0.6	4



单纯从代码行角度评价开发团队的绩效明显不可靠

群体智能和项目成功度的相关性

相关系数	评分	Watch	Star	Fork	讨论数
活跃度	0.029	0.196	0.190	0.131	0.098
开发者人数	0.123	0.368**	0.358**	0.593**	0.288*

*. 置信度95%
**. 置信度99%

开源软件的开发者人数和项目成功度强相关

软件运营过程中的数据分析实践

- 移动互联网用户数量日渐增多，
  智能手机终端逐渐成为移动媒体主要载体；

- **微信**是改变人们生活方式的重要产品之一，截止2015年12月31日，微信每月活跃用户已达到6亿。

- 基于微信平台的**运营数据分析**，可以**帮助企业更好地开展相关业务**

互联网模式下的**价值评估方式**以转向两大新的指标

- **访问流量**

微信流量不仅可以维持企业品牌知名度，还能够一定程度上转化为消费；

- **用户量**

微信用户由于其真实性和有效性，具有很大的潜在价值。

软件运营过程中的数据分析实践

数据采集

- 设计搭建微信传播分析系统，实现**传播路径检测**、**流量监测**：

  (1) 用户只需在系统上建立**监控任务**，即可便捷地搜集各种微信数据信息；

  (2) 产品在微信平台的**传播路线精确监控**，获得**用户传播点的精确数据**；

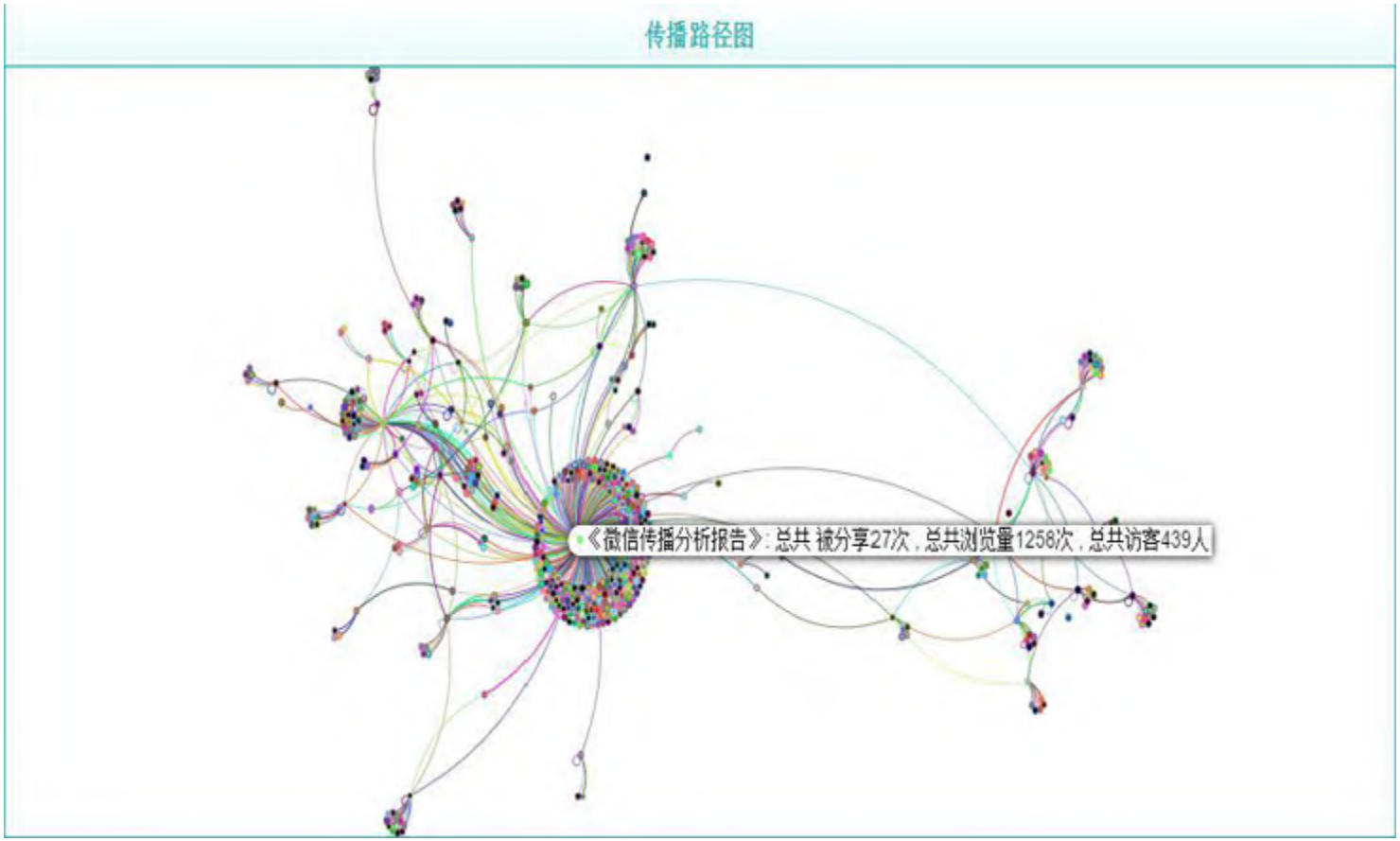
数据分析

- **挖掘关键传播**(对推广传播起到决定性影响)**用户**：

  (1) 企业采取一定的激励刺激手段，保持KOL用户的忠诚度和粘性；

  (2) 获得**潜在用户**，实现**广告精准投放等**；
- 分析用户地域信息、终端信息；
- 分析用户朋友圈画像。

传播路径监测

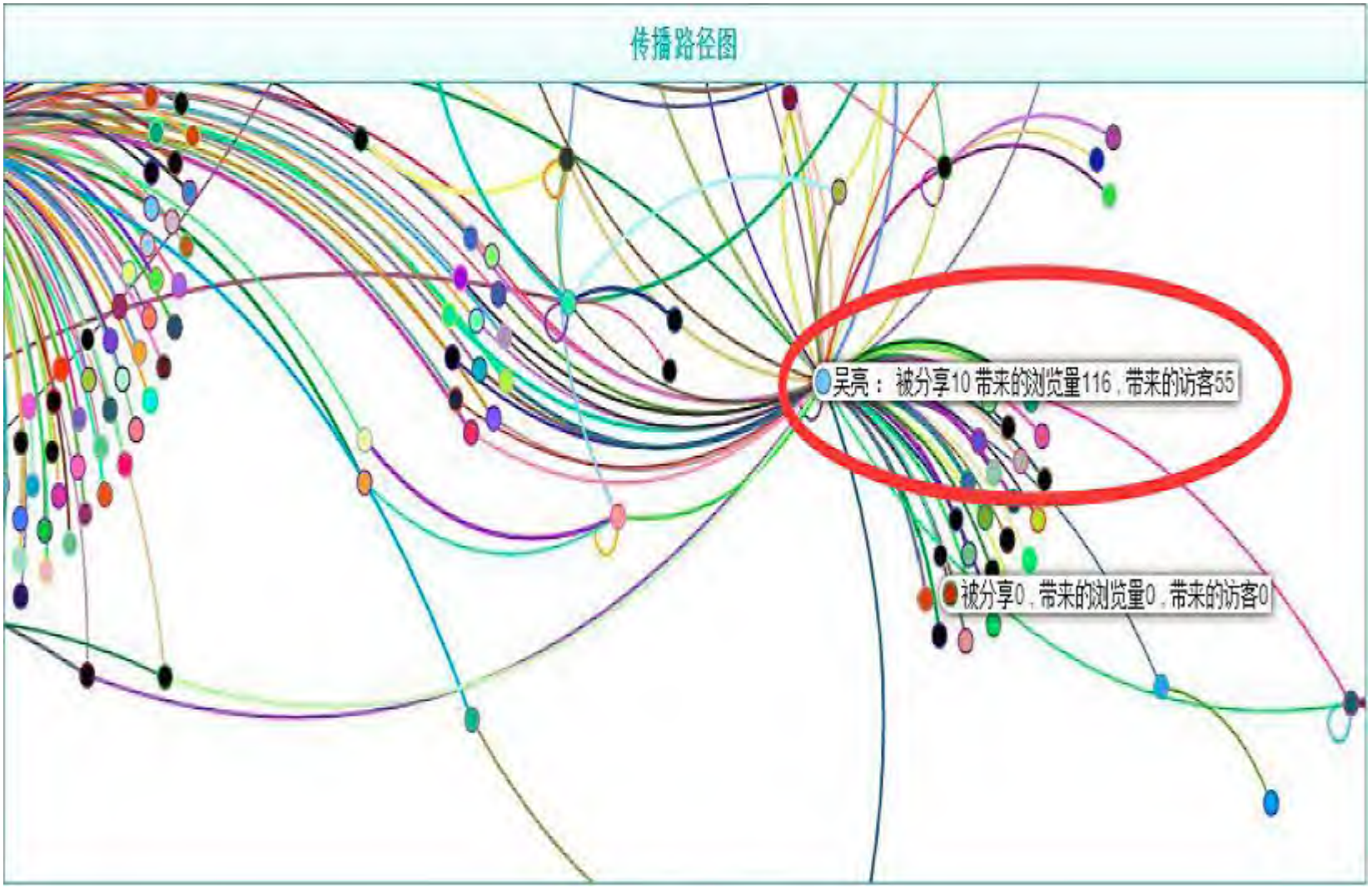


从推广源点到消息送达终点，每点代表一个用户；

每个人带来的**分享数、浏览数、实际带来的访客数**；

在圆点鼠标悬停或放大查看，即可精确获取。

关键传播用户挖掘



KOL用户挖掘案例

关键传播用户挖掘

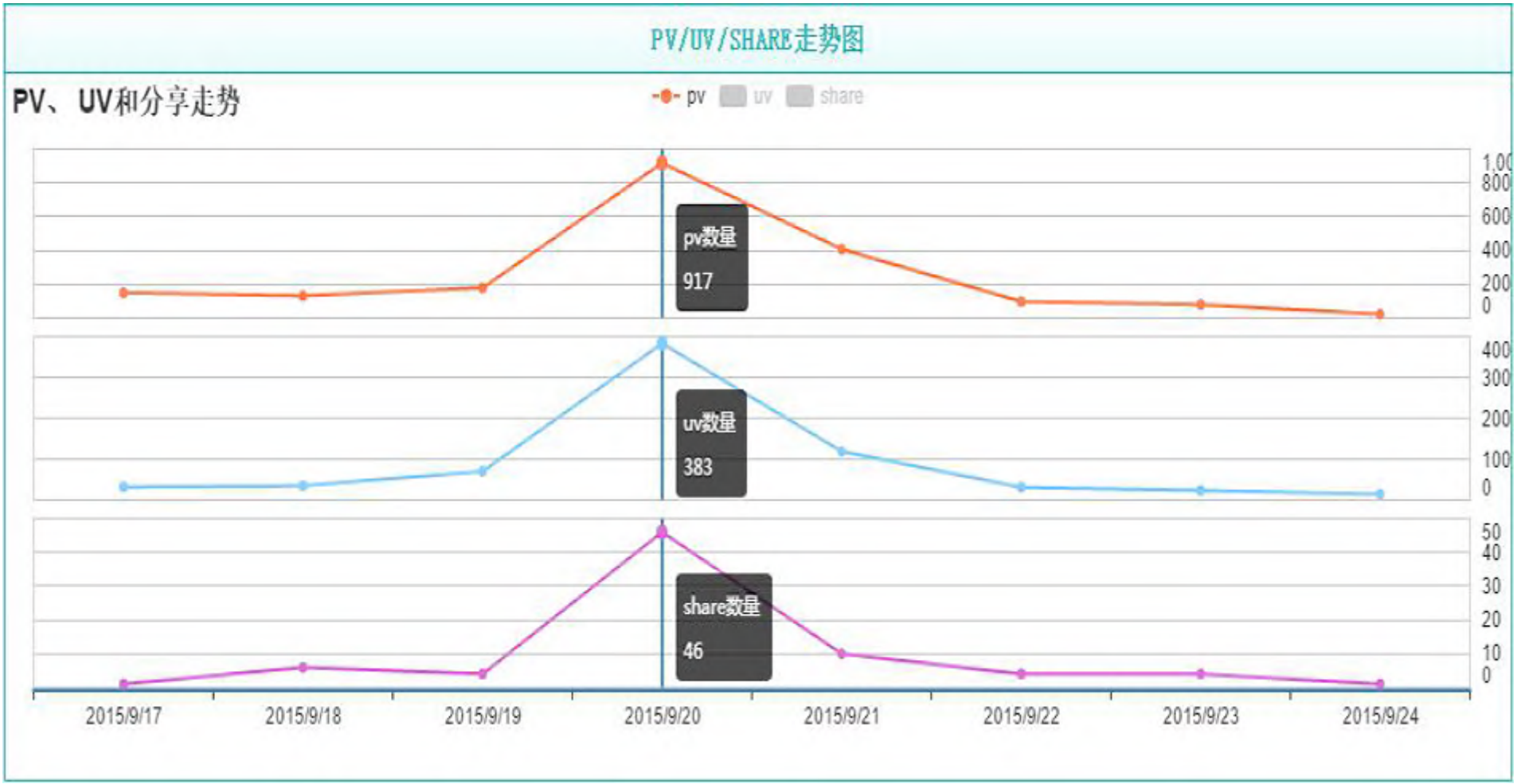
一次传播分享

10次分享、
116次浏览、
55名访客

KOL用户！！！

重点关注和维护：
1企业的活跃用户
2企业的有效潜在用户

流量监测



浏览量(PV)/访客数(UV)/IP 历史数据趋势图

流量监测

TiD2016

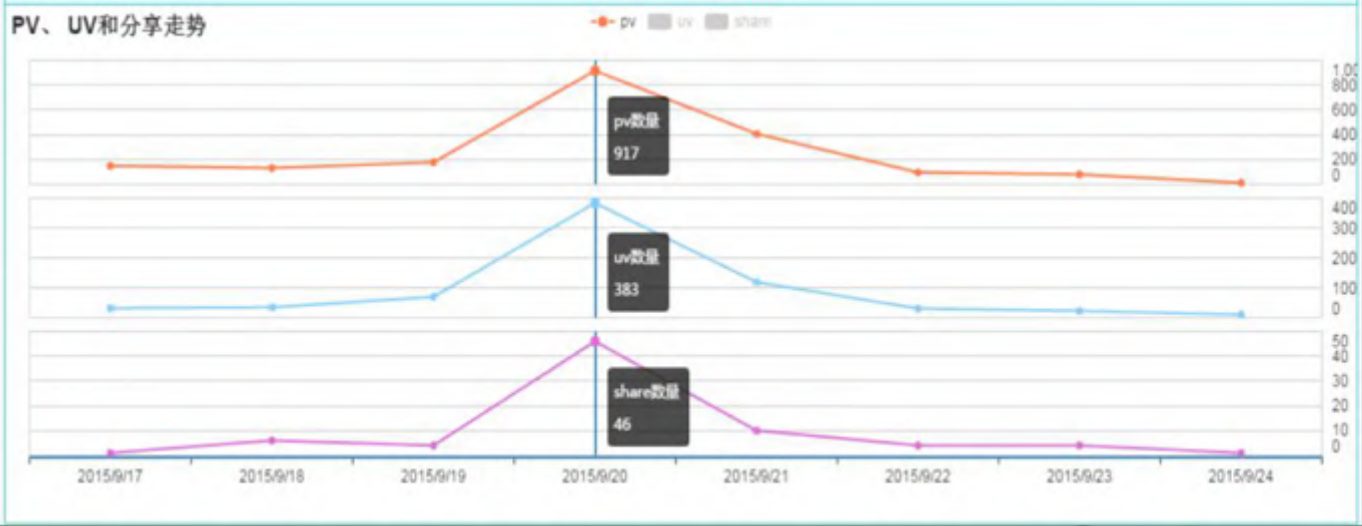
关键传播用户图表

排名	头像	昵称	性别	地址	被分享数	带来访客数	带来浏览量
1		吴亮		上海浦东新区	10	56	117
2		皮皮			5	10	29
3		蓓蓓	女	上海浦东新区	5	27	75
4		严晨秋-20KM徒步降	男	上海浦东新区	4	20	52
5		晓梦		上海卢湾	3	8	9
6		D张如琳		上海黄浦	2	4	8
7		鼎 Jing		上海浦东新区	2	11	38
8		悦余箱		上海	2	19	39
9		Ling		上海	2	10	14
10		Loki		上海	2	5	5
					下一页	第1页	共550条记录

说明

关键传播用户图表分析数据

PV/UV/SHARE走势图



说明

微信PV/UV/SHARE数据走势

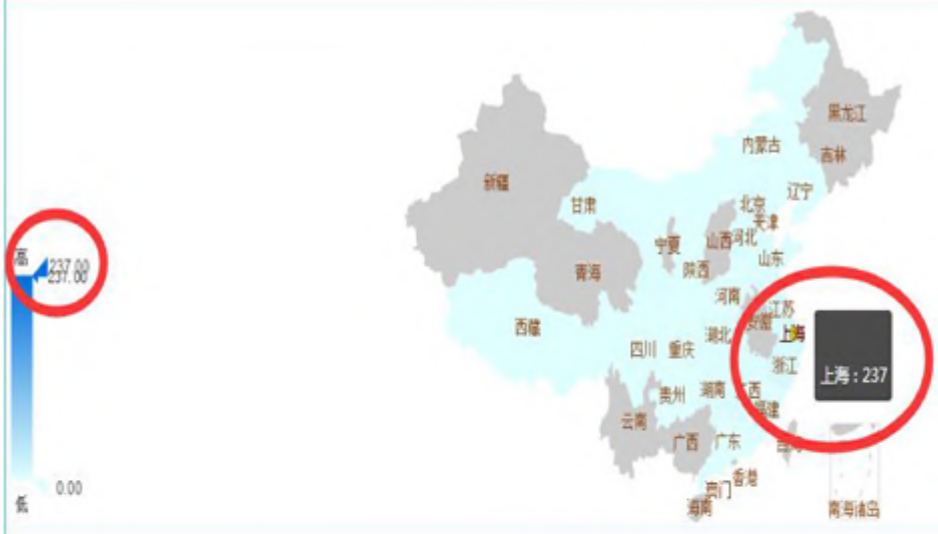
用户追踪分析

传播地域分布及方式分布

TiD2016

用户追踪分析

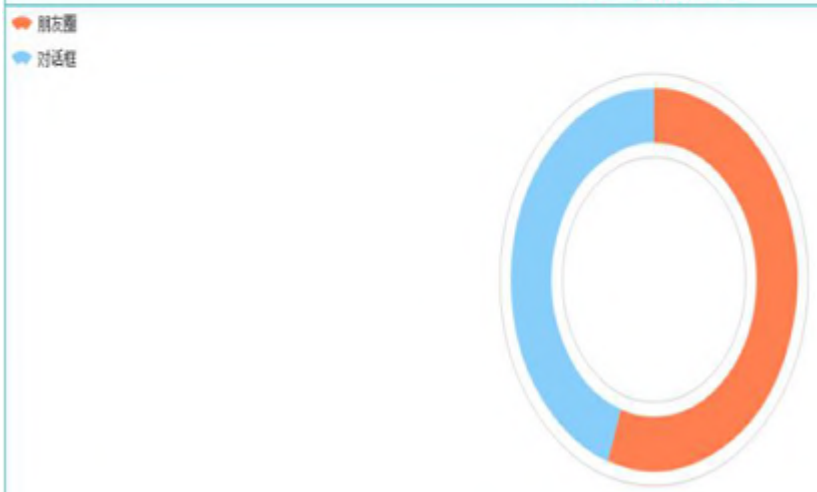
PV用户地域分布图



说明

地域分布图

传播分享去向分布图



说明

传播分享去向分布图

用户追踪分析

TiD2016

PV/UV/IP数据统计图表					说明
	浏览量(PV)	访客数(UV)	IP数	分享数	今日/昨日/总数统计数据
2015-09-22	18	10	11	0	
2015-09-21	407	119	152	10	
总数	425	129	163	10	

- 基于流量统计，挖掘信息热度的关注点，设定企业营销计划等；
- 基于地域分布，挖掘各地区信息热点，定向推广服务等；
- 基于方式分布，挖掘较为普及的信息传播方式，选择广告、服务等的推广方式。

传播内容偏好分析

Q1： 高阅读量、高转发量的帖子是其中的何种因素吸引了用户？

Q2： 营销活动的优惠最终被哪些用户所接受了？

- 传播文章：挖掘用户对于帖子内容的偏好特征

  (1) 文本建模：提取帖子内容、相关配图、版面设置等特征，文本分析挖掘模式特征和内容主题；

  (2) 用户建模：将用户基本信息（如可获取）与用户的阅读时间、阅读量结合。

- 优惠活动：与传播文章类似，提取其力度、形式等特征，挖掘用户群组的兴趣点。

研究结果

- 用户分析结果

  发现用户中的活跃的传播者

  发现新用户获取的路径或地理位置

- 内容分析结果

  文章主题及内容图片的传播影响效果

  优惠活动要素对活动参与性的影响效果

- 综合分析模型

  用于对待发送内容进行传播优化，并匹配适当目标客群，根据过往数据训练后得出的有效预测模型

研究结果 -- 微信量化运营收益

内容分析与用户分析结合：

- 更优化配置线上营销资源

(1) 在宝贵的推送机会中给用户带去更符合其个人喜好的资讯

(2) 将营销优惠通过更有效的传播路径，以适当的营销成本投放出去

- 更快速累积潜在用户

(1) 保持与关注用户间的活跃度，维持在客户心中的品牌优先级

(2) 通过老客户，获取新客户，扩大微信渠道用户量，为后期客户转换提供充足潜在客户资源

研究结果 -- 微信量化运营理念

- 基于传统市场营销及客户管理的方法
- 引入定量分析宣传内容的模式，补充定性分析方法的不足
- 传播的文本及图片将基于传播对象以量化方式被评估
- 精确有效地衡量每个现有用户对企业流量及客户获取带来的收益
- 更合理地投入活动及优惠成本，使其事半功倍

THANKS
点燃成功
点燃梦想
谢谢