TiD 2017
质量竞争力大会

JJ比赛
WWW.JJ.CN

数据挖掘中的敏捷与精益

竞技世界（北京）网络技术有限公司
巴 川

下一代软件研发
SOFTWARE DEVELOPMENT

巴川

- 竞技世界资深数据挖掘专家；
- 北航软院兼职硕导；
- 曾在中国搜索、乐视、搜狐畅游、竞技世界等公司从事互联网数据挖掘工作；
- 主要研究领域：互联网运营分析、产品分析，社交网络挖掘、推荐系统、数据可视化。

目录

01 敏捷与精益

关于敏捷

01 敏捷就是快？

02 敏捷只适用于轻量级项目？

03 敏捷团队规模该多大？

04 敏捷主要靠人or制度？

关于精益

01 精益快or慢？

02 精益开发or精益管理？

03 精益能保证一次就做对吗？

04 精益主要靠人or制度？

02 数据挖掘项目特点

01 大数据，低价值

02 精准计算与超强容错

03 模型精度与泛化能力

04 不同业务不同追求

05 数据挖掘的目的与本质

03 数据挖掘案例

1. 用户行为路径
2. 用户搜索网络
3. 用户挽留与封杀

01 用户行为路径

（一）整体云图

可以更直观地观测：

- 用户访问的主要跳转路径及变化趋势
- 用户访问的主要页面节点及变化趋势
- 异常页面及异常路径—**出乎意料的更有价值！**

eg：10月17日用户访问路径云图。


pictures_wallpaper_detail - 50888.0
apps_featured - 32408.0
downloadmanager - 193270.0
games_featured - 317
home - 209298.0
apps_detail - 75394.0
games_detail - 47328.0
emenu_appmanager_update - 24543.0
t picpal a*t picdetail

主要发现

方便找出主要节点、路径或异常节点路径。

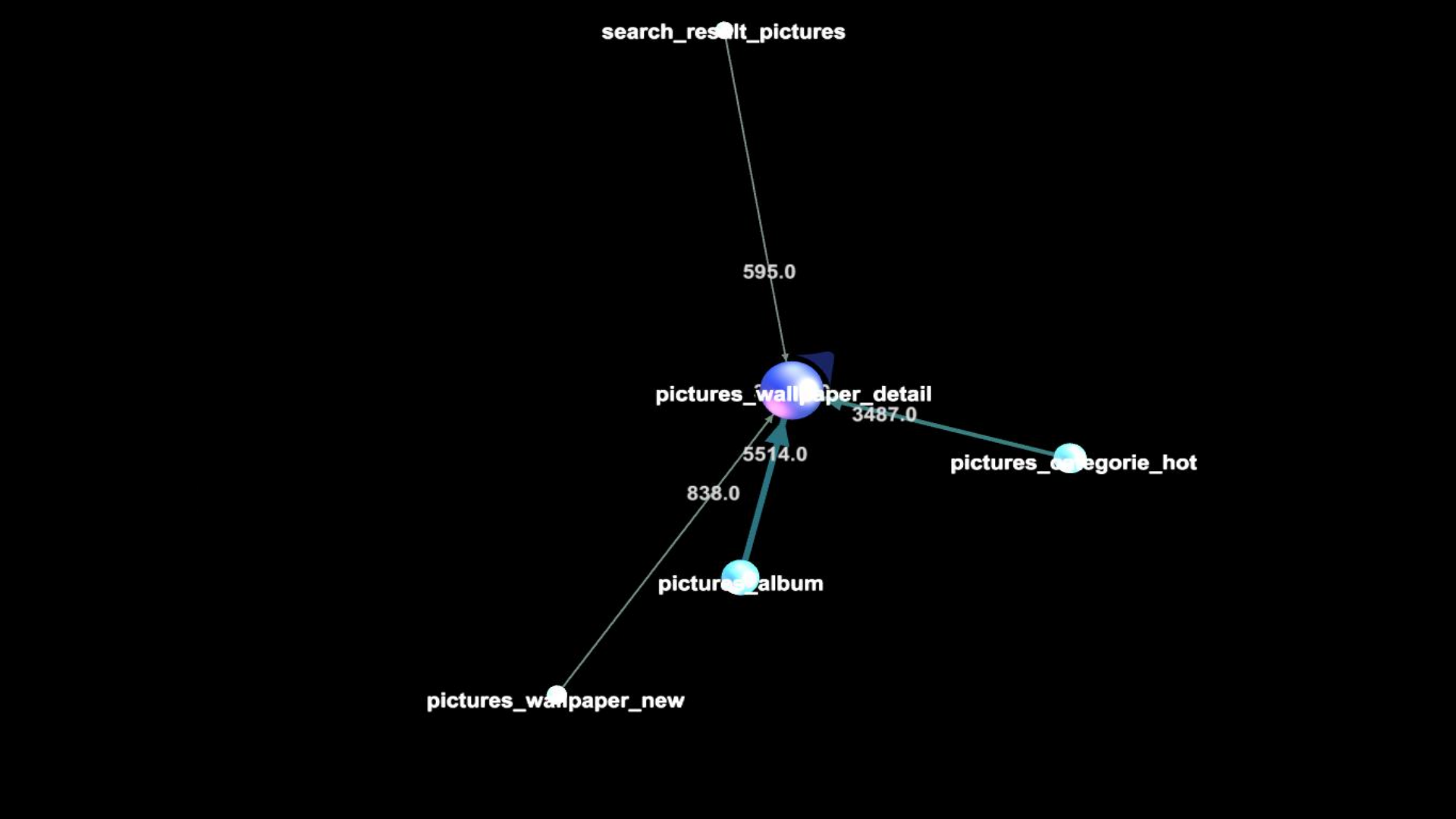
10月17日壁纸详情页访问非常高，超过app和game，因此可用行为路径云图查看其单个页面的访问来源。

（二）单个页面来源分析

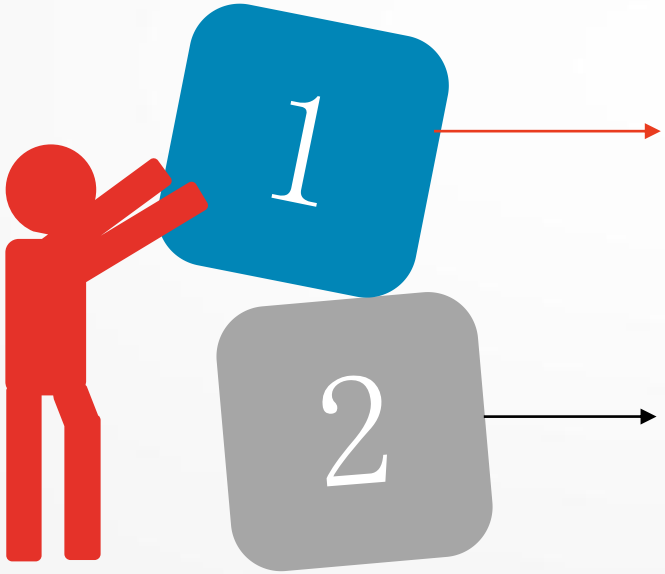
主要观测某个页面的跳转来源：

- 监测主要的来源页
- 发现异常高来源页
- 发现出乎意料低的来源页

eg：刚才异常高的壁纸详情页的跳转来源。


595.0
3487.0
5514.0
838.0

主要发现



过滤较小边之后可发现来源最多的四个页面为：

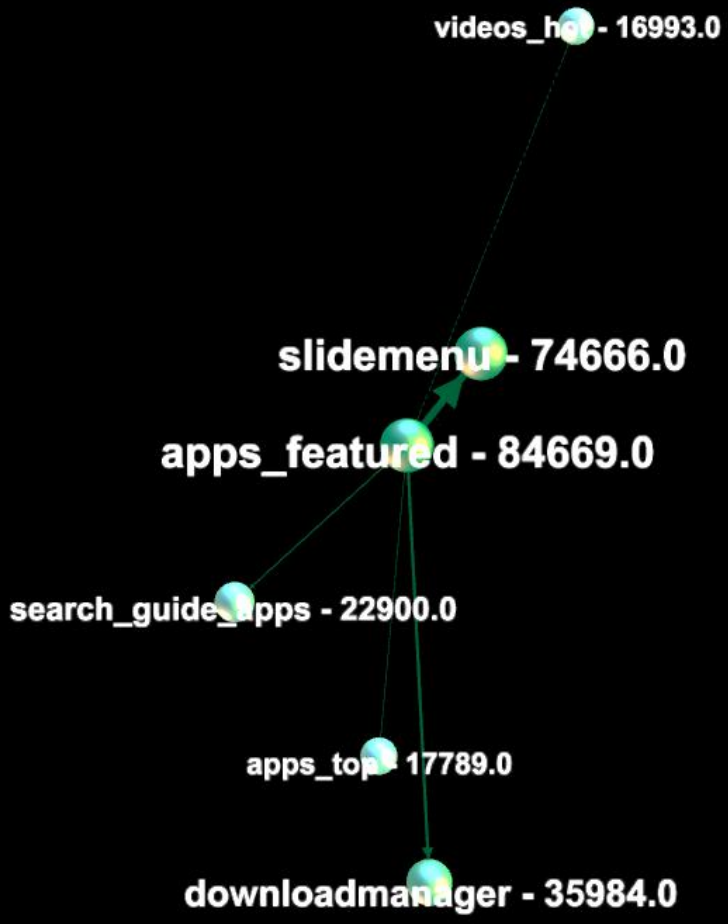
- pictures_album
- picture_categorie_hot
- picture_wallpaper_new
- search_result_pictures

除了图片资源内部跳转，从搜索过来的较多。

（三）单个页面去向分析

主要观测某个页面的跳转去向：

- 发现用户更习惯从当前页面去向哪里
- 找出与设计初衷不符的用户习惯，进行改进


videos_ho - 16993.0
slidemenu - 74666.0
apps_featured - 84669.0
search_guide_apps - 22900.0
apps_top - 17789.0
downloadmanager - 35984.0

主要发现

过滤较小边之后可发现除自身跳转外从首页去向最多的是：

- 功能类的slidemenu、downloadmanager
- 搜索类的search_guide_apps
- 资源类的app_top、video_hot

（四）主题行为分析

主要观测某个页面的跳转去向：

- 发现用户更习惯从当前页面去向哪里
- 找出与设计初衷不符的用户习惯，进行改进


music_a umdetail
music_a umlist
search_re ult_music
136.0
mus play
119.0
91.0
mus top
85.0
music_ ymusic
music_ ngtones
music_mymu c_localmusic
downloa nanager

主要发现

过滤较小点之后发现音乐模块来源最多的页面为：
Home
Downloadmanager
search_result_music

1

到达最多的页面：
music_albumlist
music_albumdetail

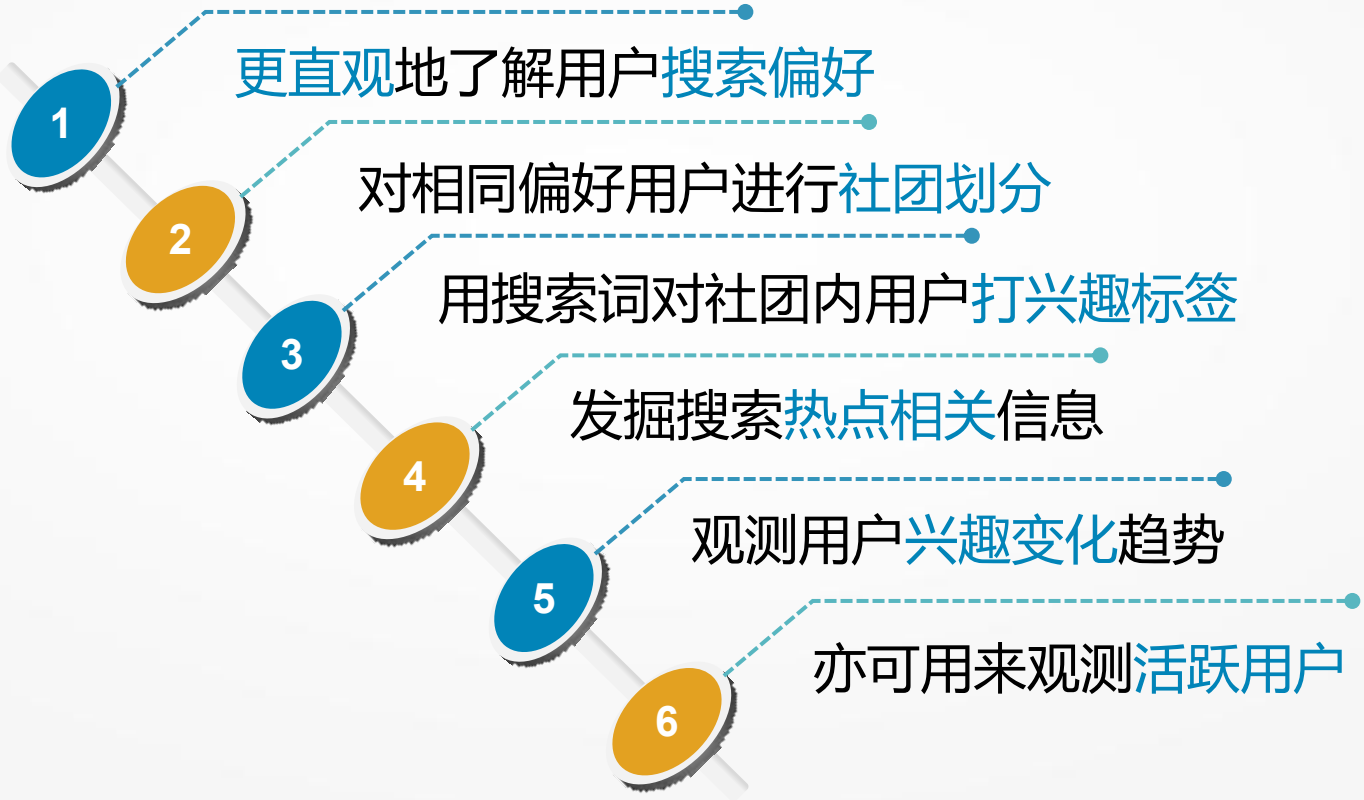
2

亦可单独分析某页面的来源和去向

3

02 用户搜索网络

用户搜索观星台


1
更直观地了解用户搜索偏好
2
对相同偏好用户进行社团划分
3
用搜索词对社团内用户打兴趣标签
4
发掘搜索热点相关信息
5
观测用户兴趣变化趋势
6
亦可用来观测活跃用户

（一）站点之星

eg：IN站10月10日用户搜索数据。


full games
antivirus security - free
gmail
amazon india
flipkart
amazon

主要发现

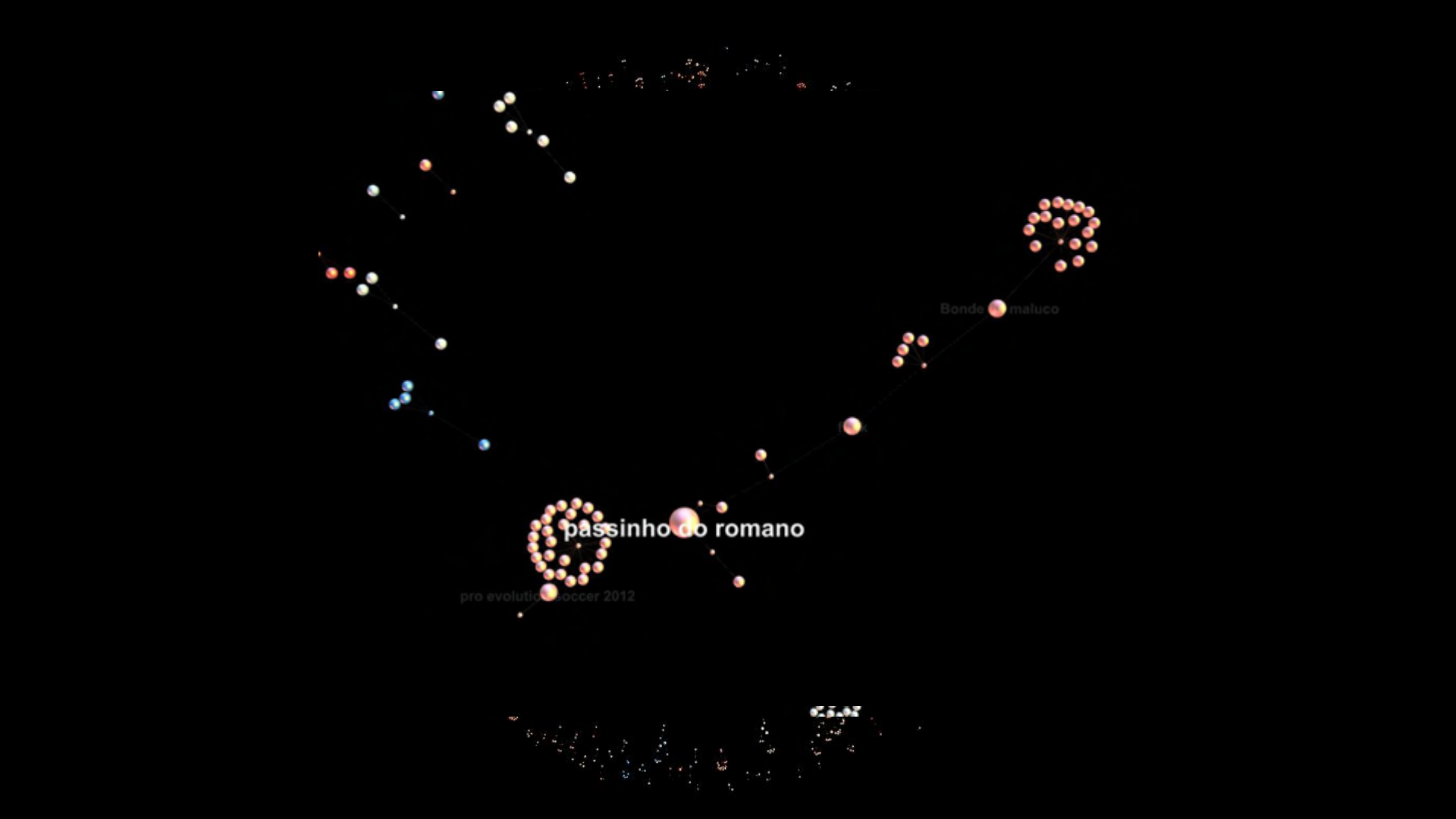
1. 在印度，Whatsapp、facebook仍然比较流行
2. 色情词依然较热：hot sex porn XXX video、sex games等
3. 视听类比较流行：youtube、mx player、tv等
4. UC浏览器在印度有一定市场：UC blowser
5. 同社团内的用户兴趣偏好趋同：图中最小节点为用户
6. 同社团内的搜索词相关性较高：如youtube与tubemate、bangbang与music player、flipkart与amazon等
7. 其他热点：flshtransfer、gta、full games

（二）斗转星移

观测用户兴趣变化趋势

数据源：BR站用户搜索数据

日期：10月4日、10月7日、10月14日


passinho do romano


livro poético
Facebook
jogos
touchpal x keyboard premium
touchpal keyboard
naruto
pes 2012
porno
pou
minecraft
minecraft - pocket edition
teclado
jogos
gangstar city
grand theft auto san andreas


sticks figure spotlight
halo 4
halo
menino e menino macaco e macaco e viado e viado
jogo do Halo
mc bim laide
mc bin laden

主要发现

色情词总是比较热门

eg：prono、xvideos

巴西人民爱足球

eg：pes2012、 pro evolution soccer2012

也爱动感音乐

eg：passinho do ramano、bonde maluco

有世界流行

eg：minecraft、facebook

也有本地流行

eg：jogo do halo

二级热点切换较快（更有价值！）

eg：4日dragon ball、7日angary bird、14日pes2012

03 用户挽留与封杀

用户流失预警


1、目标变量定义
30天不登录？
7天不登录？
30天无时长？
7天无时长？
……
2、目标用户群
日活用户？
周活用户？
月活用户？
新老用户？
零值用户？
3、时间窗确定
观察期
表现期
缓冲期
4、特征工程
基础变量：
注册时长、账号类型等
衍生变量：
均值、波幅、占比等
5、特征选择
相关变量剔除
零方差变量剔除
……
6、模型选择
逻辑回归
随机森林
SVM等等
7、模型评估
命中率
查全率
提升度
泛化能力
8、数据输出
用户名单
流失概率
相关特征

刷金用户识别

1、数据平衡

训练集：某一天日活用户＋一个月内被封的刷金用户

测试集：每天的日活用户+被封的刷金用户

2、特征工程

平台日聚合特征

游戏路径特征

比赛行为特征

3、模型选择

逻辑回归

随机森林

SVM等等

4、初期嫌疑用户

具有刷金倒金行为的刷子用户

与刷子有同样行为的正常用户

5、种子刷金用户

统计嫌疑IP

过滤掉正常用户

找到种子刷金用户

6、相似度扩展

皮尔逊相似度

硬件地址异同

7、过滤正常用户

过滤掉上一步得到用户中的如下用户：

1、手机号注册

2、邮箱注册

3、QQ微信注册

8、刷金用户名单

最终的游戏刷子名单会进入生产系统，由运营人员封杀

算法纠结与平衡

P值	命中率	查全率
0.9	90.67%	21.58%
0.8	87.79%	28.84%
0.7	83.14%	36.28%
0.6	79.56%	65.10%
0.5	**75.46%**	**76.08%**
0.4	67.79%	86.35%
0.2	55.72%	94.79%

算法纠结与平衡

P值	命中率	查全率
0.9	90.67%	21.58%
0.8	87.79%	28.84%
0.7	83.14%	36.28%
0.6	79.56%	65.10%
0.5	75.46%	76.08%
0.4	67.79%	86.35%
0.2	**55.72%**	**94.79%**

算法纠结与平衡

P值	命中率	查全率
0.9	90.67%	21.58%
0.8	87.79%	28.84%
0.7	83.14%	36.28%
0.6	79.56%	65.10%
0.5	75.46%	76.08%
0.4	67.79%	86.35%
0.2	55.72%	94.79%

小Tip

不平衡数据分类

——过采样、欠采样、SMOTE？

关于命中率和查全率

——调整分类概率阈值

有效特征最重要！

有效特征&优雅模型

“任何一个有智力的笨蛋都可以把事情搞得更大，更复杂，也更激烈。往相反的方向前进则需要一点天分，以及很大的勇气。”

–阿尔伯特•爱因斯坦

04 体会与总结

关于敏捷

敏捷就是快？	A	不只是快！
敏捷只适用于轻量级项目？	B	大项目亦可拆解
敏捷团队规模该多大？	C	“三三制”与“地泽二十四”
敏捷主要靠人or制度？	D	主要靠人！

关于精益

精益到底快or慢？	A	节奏更重要!
精益开发or精益管理？	B	敏捷开发，精益管理
精益能保证一次性做对么？	C	不能，但能少走弯路
精益主要靠人or制度？	D	主要靠人！

Thanks