案例库分析

背景

有了案例后，我们通过分析案例相关的数据定位问题，但目前数据的提取和分析的工作量很大，并且不是很方便

案例分析工具将为大家提供已经取好的数据源和基础的统计分析功能，能让大家快速定位到关键数据维度

展示示例

对比案例 大盘数据 1 2

分析案例 3 4

快速分析 高级分析

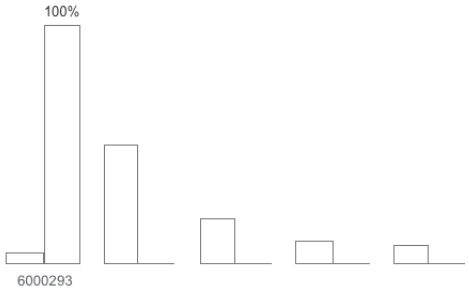
build platform ip城市 数美model netType

一阶 二阶

字段	指标	
build	9.8	高
ip城市	8.4	高
platform	4.4	高
数美model		
netType		

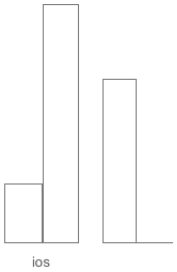
build 对比 KL散度（相对熵）：9.8

	值	对比案例		分析案例	
		数值	比例	数值	比例
1	6000293	1	<1%	31	100%
2	6920189	58	29%	0	0%
3	6920205	39	20%	0	0%
4	6930175	29	15%	0	0%
5	6930196	26	13%	0	0%
总量		200		31	



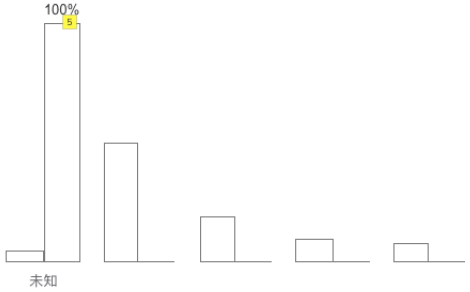
plarform 对比　　KL散度（相对熵）：4.4

	值	对比案例		分析案例	
		数值	比例	数值	比例
1	ios	112	56%	31	100%
2	android	88	44%	0	0%
总量		200		31	



ip城市 对比　　KL散度（相对熵）：8.4

	值	对比案例		分析案例	
		数值	比例	数值	比例
1	未知	1	<1%	31	100%
2	中国广东	39	20%	0	0%
3	中国浙江	24	12%	0	0%
4	中国河南	14	7%	0	0%
5	中国江苏	14	7%	0	0%
总量		200		31	



快速分析页面

分析配置组键

对比组：默认为大盘抽样

分析案例：访问后端获得可分析的案例库

分析维度：支持【单维度分析】，【双维度分析】和【三维度分析】

高级分析功能：先置为灰色不可用

分析维度选择

可选维度顺序和显示：

- 第一行： build platform ip城市 数美Model netType 设备品牌 发布数 粉丝数
- 下拉选项： 是否为国内手机号 是否近期注册登录 最近活跃省份 最近活跃国家 设备文件存储路径 手机号前7位 注册发布间隔 ip C段 模拟设备 积分墙设备 多开设备 改机设备 云控设备 农场设备 伪造设备 root机 虚拟机 切换设备数 同设备用户数 用户五分钟内发布数 手机号1日内发布数 设备1日内发布数 ip下7日内用户数 7日内相似笔记数 用户1日内高风险笔记数

更多选项按钮：

- 放在第一行的最后
- 只有一行显示时，作为【更多选项】功能；显示下拉后，作为【收起】功能

全选和反选：

- 默认的选项是第一行全选
- 下拉选项后，还是保持只有第一行全选的选项，收起时默认其他选项不选

概览栏

显示维度：显示top30 metric的维度

联动：点击维度，可以跳转到相应的具体数据框（功能开发可选）

界面显示

页面一：快速分析

表格数据选取，和doris交互

Doris取数逻辑

```
--
--
----group_id=0
----group_id=270001
----column_name=platform

select
        a.column_value,
        a.duibi,
        a.duibi/b.duibi as duibi_pt,
        a.fenxi,
        a.fenxi/b.fenxi as fenxi_pt,
        b.fenxi fenxi_cnt,
        b.duibi duibi_cnt
from
        (
                select
                        column_value,
                        sum(if(group_id=0,1,0)) duibi,                 --group_id
                        sum(if(group_id=270001,1,0)) fenxi, --group_id
                        count(1) as count,
                        column_name
                from reddw.xpander_note_data
                where (group_id = 0 or group_id=270001)
                and column_name='platform'
                group by column_name, column_value
        ) a
join
        (
                select
                        sum(if(group_id=0,1,0)) duibi,
                        sum(if(group_id=270001,1,0)) fenxi
                        from reddw.xpander_note_data
                where (group_id = 0 or group_id=270001)
                and column_name='platform'
) b
on 1=1
order by a.duibi/b.duibi+a.fenxi/b.fenxi
desc;
```

指标计算方法

KL散度

$$D_{KL}(A||B) = \int a(x) log\left(\frac{a(x)}{b(x)}\right)$$

A是对比组，B是分析组

接口

- 通过关键词查询案例库名称接口
- 给定分析字段和组合方式反馈分析结果的接口
    - 输入 column_descriptions:["build","ip城市","platform"], combine_type：["一阶"]

分析接口返回格式

```
[
        {
                column_name:"build",  --
                column_description:"build",  --description
                column_metrics:
                        {
                                kl_value:9.8,   --kl
                                kl_level:""   -- kl
                        }
                column_distribution:
                        [
                                {
                                        --build=6000293
                                        column_value:6000293,
                                        _cnt:1,
                                        _pt:"<1%",
                                        _pt_level:"",
                                        _cnt:31,
                                        _pt:"100%",
                                        _pt_level:"",
                                },
                                {
                                        column_value:6920189,
                                        _cnt:58,
                                        _pt:"29%",
                                        _pt_level:"",
                                        _cnt:0,
                                        _pt:"<1%",
                                        _pt_level:"",
                                },
                                {}
                        ]

        },
        {
                column_name:"platform",
                ...
        }
]
```

方案讨论

1. 案例库案例更新 触发 案例数据更新
    a. 第一期，我们只取少量数据维度
    b. 后期需求：维度可扩充，尽量无schema化
2. 案例分析数据展示
    a. 前后端配合

数据更新数据源

1. hammurabi中收集到的发布时的数据源
    a. tidb数据
        i. 优点：查询快速
        ii. 缺点：只有7天的数据
    b. hive数据源
        i. 优点：数据全
        ii. 缺点：查询缓慢
2. 自建数据
    a. 根据insight用户数据+设备数据拼接成数据源
        i. 优点：数据灵活性高
        ii. 缺点：没有发布当下的数据（累计、ip及其他查询）