榜单数据源推送需求

# 需求来源

需求人：李彦杰/00216639

需求部门：用户经营产品经理管理部

需求类型：数据源推送到应用市场和游戏中心

# 业务背景描述

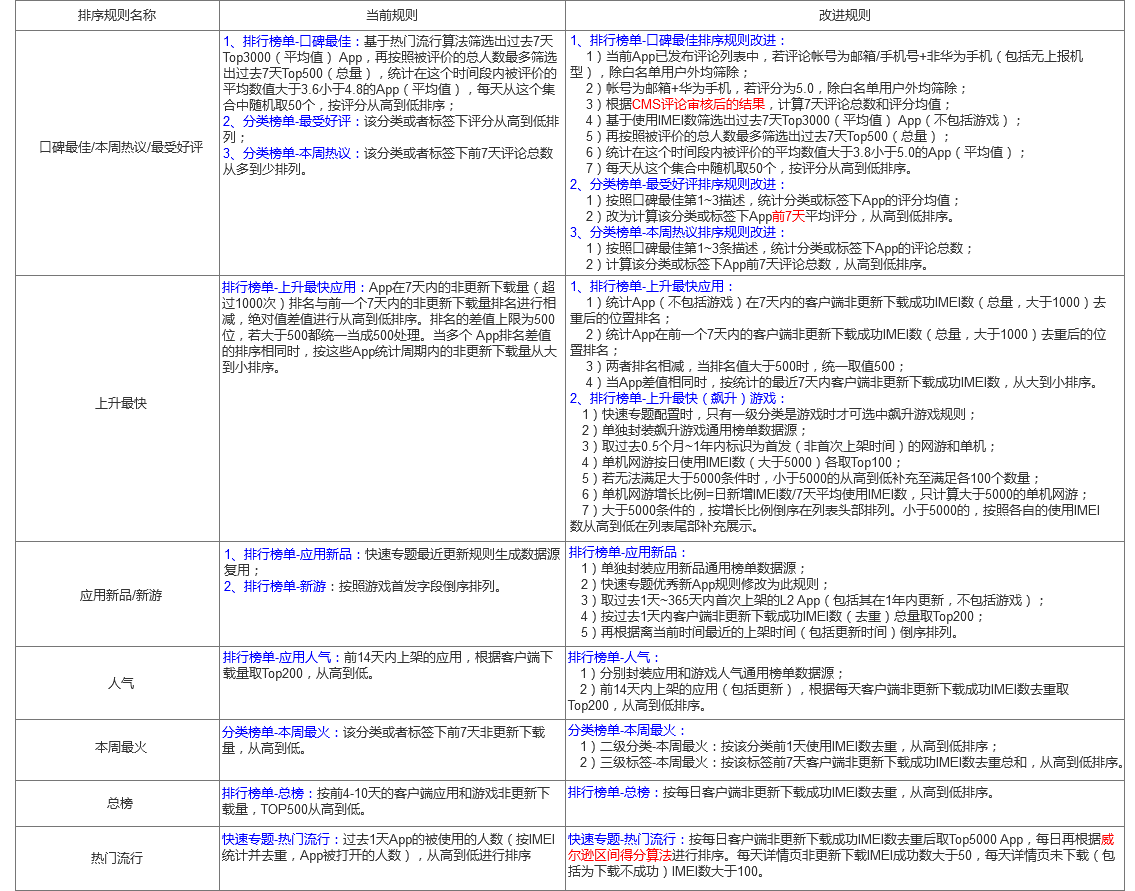
## 业务背景

应用市场SR00052QBT榜单数据源排序规则及配置改进需求，主要目的解决当前专题、排行及分类榜单排序规则不合理的问题，保持数据源一致性，从规避刷量的角度进行榜单排序。同时，针对热门流行的排序规则进行整改，采取威尔逊区间得分算法进行计算并排序。所有改进规则需要BI计算后推送给应用市场和游戏中心。

# 大数据需求描述

### 需求描述

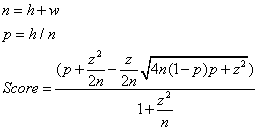
应用市场和游戏中心待改进的排序规则如下：



基于上表改进规划描述，需要BI推送以下数据源：

1. 客户端非更新下载成功的IMEI数（前1天、去重、HOTA关联、所有L2 App）；
2. 客户端非更新下载成功的IMEI数（前7天、去重、HOTA关联、所有L2 App）；
3. App日使用IMEI数（前1天、去重、HOTA关联、所有L2 App）；
4. 威尔逊区间得分排序（前1天、应用和游戏各Top200）。

威尔逊区间得分计算公式：



1. 正式名称为：二项分布样本的威尔逊置信区间下限，计算公式是在1927年由数学

家 Edwin Bidwell Wilson 提出的，用来对二项分布进行参数估计。一般用威尔逊置信区间的下限对具有正负双向投票的系统进行排序，目前是Reddit网站Best帖子的排序算法；

1. p越大，就代表这个项目的好评比例越高，越应该排在前面。但是，p的可信性，

取决于有多少人投票，如果样本太小，p就不可信。，p是"二项分布"中某个事件的发生概率，因此可以计算出p的置信区间。所谓“置信区间”，就是说，以某个概率而言，p会落在的那个区间。比如，某个产品的好评率是80%，但是这个值不一定可信。根据统计学只能说，有95%的把握可以断定，好评率在75%到85%之间，即置信区间是[75%, 85%]。所以该算法的思路如下：

第一步，计算每个项目的“好评率”（即赞成票的比例）。

第二步，计算每个"好评率"的置信区间（以95%的概率）。

第三步，根据置信区间的下限值，进行排名。这个值越大，排名就越高；

置信区间的实质，就是进行可信度的修正，弥补样本量过小的影响。如果样本多，就

说明比较可信，不需要很大的修正，所以置信区间会比较窄，下限值会比较大；如果样本少，就说明不一定可信，必须进行较大的修正，所以置信区间会比较宽，下限值会比较小，而威尔逊公式很好地解决了小样本准确性问题；

1. 现有的总榜单、人气榜、新秀榜单均按照非更新下载量组合排序，过于简单。通过

此算法的概念，可以将详情页非更新下载成功IMEI数，来作为一个App非更新下载的热度样本，最终虽然是按非更新下载成功排序，但并不是下载量最高的就会排在头部，而是对于用户真正有吸引力的在前面；

4、参数：

1）h为进入到某App详情页并下载成功此App的IMEI数去重加权，应用市场和游

戏中心统计大于50的App；

2）w为进入到某App详情页未下载（包括为下载不成功）此App的IMEI数去重

加权，应用市场和游戏中心统计大于100的App；

3）z为可调常数，即置信区间大小，一般情况下，在95%的置信水平下，z统计量

的值为1.96；

1. Score的范围为（0，1），基于热度的置信度二项式分布公式，若n很大，表示该

App有热度，同时p也很大趋近1时，该公式得分趋近于1，Score最重要的影响因素是p，即下载成功的比值。若n很小，表示该App冷门，关注度小，若同时p很小时，该公式得分趋近于0。

试算结果：

### 低保真设计

无

### 需求场景流程说明

BI推送计算结果给应用市场和游戏中心，应用市场和游戏中心获取数据后，根据业务规则进行二次计算和重组织生成排序列表。

### 关键指标说明

|  |  |  |  |
| --- | --- | --- | --- |
| **关键指标说明** | | | |
| **序号** | **指标** | **指标说明** | **算法、统计口径** |
| 1 | 客户端非更新下载成功的IMEI数 | 返回数据格式为：  AppID|前1天（前7天）客户端非更新下载成功IMEI数 | 1、推送所有L2等级App，前1天HOTA关联的客户端非更新下载成功的IMEI去重；  2、推送所有L2等级App，前7天HOTA关联的客户端非更新下载成功的IMEI去重。 |
| 2 | App日使用IMEI数 | 返回数据格式为：  AppID|App日使用IMEI数 | 推送所有L2等级App的使用量，前1天HOTA关联的IMEI去重； |
| 3 | 威尔逊区间得分排序 | 返回数据格式为：  AppID|威尔逊得分 | 1、根据威尔逊公式进行计算；  2、 h为进入到某App详情页并下载成功此App的IMEI数去重加权，应用市场和游  戏中心统计大于50的App；  3、w为进入到某App详情页未下载（包括为下载不成功）此App的IMEI数去重  加权，应用市场和游戏中心统计大于100的App；  4、z暂定为1.96；  5、进详情IMEI数去重，统计ods\_hispace\_oper\_log\_dm中oper\_type 1~10的记录数，imei去重；  6、进详情非更新下载成功该App的IMEI数去重（客户端），统计ods\_hispace\_down\_install\_log\_dm中optype=1 and result=0 and info rlike ‘detail:1’ and !(info6 rlike ‘renew’)  imei去重；  7、进详情未下载或非更新下载未成功该App的IMEI数去重（客户端），统计  第一列 – 第二列 – 详情页更新  详情页更新的口径：ods\_hispace\_down\_install\_log\_dm中optype=1 and result!=0 and info rlike ‘detail:1’ and info6 rlike ‘renew’  imei去重；  8、客户端详情页非更新下载量，统计  ods\_hispace\_down\_install\_log\_dm中optype=1 and info rlike ‘detail:1’  and !(info6 rlike ‘renew’) |

### 数据源依赖说明

无。

# 其他要求（备注）

# 附件