|  |  |  |  |
| --- | --- | --- | --- |
| 文件编号 | 2012-RDD-DM-0004 | 作　　者 | 苏文魁 |
| 文档版本 | V2.098E | 最后修改日期 | 2012-12-11 |

**版本号V2.098E**

**RTB实时竞价算法说明书2.0**

**编 写 人： 苏文魁**

**编写时间： 2012年8月**

**修订控制页**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **编号** | **文档版本** | **修订章节** | **修订原因** | **修订日期** | **修订人** |
| 1 | V2.091 | 第3章 | 完善竞价数据统计方法 | 2012-09-05 | 王曼 |
| 2 | V2.092 |  | 更改几个重要名词，细化统计方法 | 2012-09-10 | 王曼  苏文魁 |
| 3 | V2.093 | 名词章节提前，更改第2章，第3章细节 | 讨论业务需求和实现方法后完善 | 2012-09-25 | 王曼，苏文魁 |
| 4 | V2.094 | 增加6.5节 | 广告集合竞价 | 2012-09-29 | 苏文魁 |
| 5 | V2.095 | 增加系统需求，细化数据需求，  增加竞价指数计算公式 | 完善竞价框架，增加系统需求和性能指标需求 | 2012-10-10 | 苏文魁，王曼 |
| 6 | V2.096 | 增加附录A 数据输入输出详细字段 | 明确各系统接口需求 | 2012-10-18 | 苏文魁 |
| 7 | V2.097 | 增加实时竞价列表及使用，完善竞价流程，用竞价指数替代原来打分方式，并增加系统内广告活动竞价价格计算方法 | 业务需求 | 2012-10-21 | 苏文魁 |
| 8 | V2.098 | 修改在线竞价流程，增加实时竞价列表使用 | 业务需求 | 2012-10-23 | 王曼 |
| 9 | V2.098A | 更改竞价指数中预算指数计算公式并限定取值范围 | 数据需求 | 2012-11-9 | 苏文魁 |
| 10 | V2.098B | 增加算法修订章节：附件B1~4 | 算法适应实际数据情况调整 | 2012-11-28 | 苏文魁 |
| 11 | V2.098C | 增加是否首屏显示参数 | 提高算法效果 | 2012-12-03 | 苏文魁 |
| 12 | V2.098D | 修正DMP统计数据和分析数据表A1.2.2中活动维度表，A.3.2竞价分析 | 修正原有实现与设计不统一部分 | 2012-12-10 | 苏文魁 |
| 13 | V2.098E | 增加点击率预测数据抽样 | 提高算法效果 | 2012-12-11 | 王曼 |
| 14 | V2.098F | 完善点击率预测公式（B7） | 提高预测稳定性及精度 | 2012-12-18 | 苏文魁，王曼 |

目录

[名词 5](#_Toc338777314)

[1. RTB竞价算法 6](#_Toc338777315)

[2. RTB 关键问题分析 6](#_Toc338777316)

[2.1 RTB竞价的关键阶段 6](#_Toc338777317)

[2.2 广告筛选的主要因素 6](#_Toc338777318)

[2.3 影响广告价格的关键因素 6](#_Toc338777319)

[2.4 竞价广告选择的关键因素 7](#_Toc338777320)

[2.5 RTB系统性能需求 7](#_Toc338777321)

[3. 竞价价格（BID）预测 7](#_Toc338777322)

[3.1 总体框架图 7](#_Toc338777323)

[3.2 总体流程图 8](#_Toc338777324)

[3.3 属性选取及统计编码方法 9](#_Toc338777325)

[3.2.1属性选取 9](#_Toc338777326)

[3.2.2属性编码 10](#_Toc338777327)

[3.4 离线部分 11](#_Toc338777328)

[3.3.1 构建子样本编码库 11](#_Toc338777329)

[3.3.2 子样本竞价分布预测 12](#_Toc338777330)

[3.3.3 活动竞价价格预测 12](#_Toc338777331)

[3.5 在线部分 13](#_Toc338777332)

[3.6 在线竞价记录及使用 14](#_Toc338777333)

[3.7 竞价价格评估与优化 15](#_Toc338777334)

[4. 转化率（eTR）预测 15](#_Toc338777335)

[4.1 离线部分 15](#_Toc338777336)

[4.2 在线部分 15](#_Toc338777337)

[5. 用户价值估值体系 15](#_Toc338777338)

[5.1 基础价值（UV\_BASE） 16](#_Toc338777339)

[5.2 行为价值(UV\_BT) 17](#_Toc338777340)

[5.3 竞价价值(UV\_BID) 18](#_Toc338777341)

[5.4 估值体系评估与优化 18](#_Toc338777342)

[6. 竞价 19](#_Toc338777343)

[6.1 指数计算 19](#_Toc338777344)

[6.2 竞价方法 19](#_Toc338777345)

[7. 竞价广告的选择策略[此章由第六章替代] 20](#_Toc338777346)

[7.1 策略选择优化 20](#_Toc338777347)

[7.2 广告池优化 20](#_Toc338777348)

[7.3 用户估值投票 20](#_Toc338777349)

[7.4 总体选择优化方案 21](#_Toc338777350)

[7.5 广告集合竞价 21](#_Toc338777351)

[8. 数据输入输出及系统需求 21](#_Toc338777352)

[8.1 数据输入 21](#_Toc338777353)

[8.1.1 DMP数据需求 21](#_Toc338777354)

[8.1.2 DSP数据需求 21](#_Toc338777355)

[8.1.3 竞价数据需求 22](#_Toc338777356)

[8.1.4 系统需求 22](#_Toc338777357)

[8.2 数据输出 22](#_Toc338777358)

[8.2.1 竞价输出 22](#_Toc338777359)

[8.2.2 算法效果验证输出 22](#_Toc338777360)

[8.2.3 活动目标修正统计 23](#_Toc338777361)

[9. 参考文献 24](#_Toc338777362)

[附录A 数据输入输出详细字段 25](#_Toc338777363)

[A.1 DMP数据需求 25](#_Toc338777364)

[A.1.1用户属性数据 25](#_Toc338777365)

[A.1.2竞价数据 26](#_Toc338777366)

[A.2 DSP数据需求 27](#_Toc338777367)

[A.2.1 DSP活动表 27](#_Toc338777368)

[A.2.2渠道设置表 28](#_Toc338777369)

[A.3竞价数据 28](#_Toc338777370)

[A.3.1 流水数据 28](#_Toc338777371)

[A.3.2竞价分析数据表 28](#_Toc338777372)

[A.4其它 29](#_Toc338777373)

[A.4.1媒体维度默认值表 29](#_Toc338777374)

[A.4.2广告维度默认值表 29](#_Toc338777375)

## 名词

1. eTR（expected Through Rate）：期望转化率
2. ePA(expected Per Action):期望每次行为价格（行为包含Impression（展现），Click（点击），Sale（售卖），Register（注册）等）
3. UV（User Value）：用户价值
4. I（Impression）：展现
5. C（Click）：点击
6. N（coNversion）：转化
7. BID：竞价价格
8. UV\_BASE：用户基础价值
9. UV\_BT：用户行为价值
10. UV\_BID：用户竞价价值[+向广告主收的钱]
11. 样本：竞价流水数据的每一条数据称为一个样本
12. 子样本：每个属性不可拆分的样本子集
13. AdValue：广告价值，本次广告成功会给广告主带来的价值
14. AdCreativeType:广告创意类型，包括文本，图片，Flash，Video等。
15. AdMaxBid：广告活动的最高出价。
16. AdServiceType：广告投放类型，目前包括CPM,CPC。
17. eSucRatio：预期成功率
18. CPI（Campaign Price Index）：竞价指数
19. RANK：广告出价得分
20. ：RANK的初始值
21. ：回头客分值
22. ：新客分值
23. ：商品关联分值。
24. ：单位价值到RANK的转化系数
25. ：该时段期望展示的次数
26. ：该时段实际展示的次数
27. ：展示在打分中占的分值
28. ：用户估值在打分中所占分值
29. BugetI：预算消耗指数
30. AdPrice：收取广告主价格

## RTB竞价算法

RTB（Real-Time Bidding）实时竞价，是一种利用第三方技术在数以百万计的媒体上针对每一个用户展示行为进行评估以及出价的竞价技术。与大量购买投放频次不同，实时竞价规避了无效的受众到达，只针对有意义的用户进行购买。它的核心是DSP平台（需求方平台），在DMP数据的支持下根据媒体，广告特点和人的属性进行定向投放。RTB对于媒体来说，可以带来更多的广告销量、实现销售过程自动化及减低各项费用的支出。而对于广告商和代理公司来说，最直接的好处就是提高了效果与投资回报率。

RTB算法有几种常见的策略：展现优化（针对品牌推广）目标，点击率（CTR）优化目标和ROI(投资回报率)优化目标。其核心都是要做到合适的广告以合适的价格展现给合适的人，找到新客，找回老客。

## RTB 关键问题分析

### RTB竞价的关键阶段

在时间上，可以将RTB竞价算法分成如下四个关键的阶段：

1. 广告筛选

该部分根据广告位的信息与系统中广告集合进行条件匹配，选择出适合该广告位的广告（集）。

1. 用户估值

根据用户的基础价值和行为价值，结合用户与广告的匹配关系估算出用户在本次竞价中对该只广告表现出的价值。

1. 竞价

以竞价成功为目标，给出每只广告的价格（通过修改活动的竞价价格和该活动对应的平台出价，调整竞价指数）。

1. 出价

根据系统整体的盈利模式及商业策略等因素，选择出一只合适的广告参与竞价。

### 广告筛选的主要因素

1. 时间
2. 频次
3. 广告创意规则
4. 产品形态定向
5. 媒体
6. 地域
7. 广告活动状态（有效以外的状态）
8. 结合DMP和AdEx媒体信息做人群属性筛选

广告筛选涉及到更多的是业务逻辑规则，筛选方案由业务部门给出，本文不做详细描述。

### 影响广告价格的关键因素

影响一只广告的竞价价格的因素很多，总体上可以分为如下几个方面：

1）用户属性

2）媒体属性

3）广告(主)属性

4）市场的竞争程度(包含渠道因素)，广告的预算情况等

### 竞价广告选择的关键因素

在广告筛选后形成的广告集合中选择合适的广告参与竞价，其影响因素总体上可以概括如下：

1. 用户价值（广告和用户的结合度最大化）
2. 系统的盈利模式（用户满意度最大化和平台收益最大化）
3. 广告活动价值
4. 系统整体广告的均衡显示策略等

### RTB系统性能需求

1. 海量数据离线分析能力；
2. 大数据在线实时采集，收集，计算，分析能力。
3. 系统应具有流量实时切换功能，以方便验证算法并分版本上线。

## 竞价价格（BID）预测

竞价价格是由市场上该广告的价格决定的，所以广告市场价格的直接来源是竞价的历史数据。所以最可靠的方法是根据竞价历史情况，在合适的维度上对广告价格进行统计，并根据预期的成功率给出广告的竞价价格。

要做到这一点要完成两个重要的事情：

1. 离线时在哪个维度上进行统计；
2. 在实时竞价时如何快速的查询历史并推断竞价价格。

### 总体框架图



图3.1 RTB竞价算法总体框架图

2.1节中描述了RTB竞价的四个关键阶段，其中广告筛选主要是根据广告的尺寸，面向人群等信息进行指标层次的选择，不涉及到智能算法，故在本文不做详细描述。本文主要关注用户的价值评估、每只广告活动在本次实时竞价中的出价和在出价活动集合中最终选取哪只广告参与竞价三个问题。

通过在竞价历史采用优化的编码方法基础上进行统计，并参照给定的成功率预测给出一只广告的竞价出价，如果该竞价条件在实时竞价记录中可以找到，采取实时竞价方法给出价格。

本文引进一种在用户估值基础上的打分方法，利用打分在出价活动集合中选择哪个广告活动最终参与竞价。

特别的，对于一个智能分析基础上的DMP系统，对于系统内所收集的用户的估值不仅对于实时竞价环境，而且对于其他广告相关系统也十分必要。本文提出一种基于用户历史信息和行为的用户估值体系，并以此估值体系为基础给出在实时竞价环境下用户估值的验证手段和使用方法。

[如果无历史数据/ DMP无此用户情况下，竞价采用默认出价，选择时采用最高出价；或选择不出价。如采用默认出价，价格需要根据行业规则制定]

[\*需要设计一种简单的编码与调整机制，来处理数据缺少或统计模型失败的情况]

### 总体流程图



图3.2总体流程图

### 属性选取及统计编码方法

#### 3.3.1属性选取

属性选取是指从一组已知的属性集合中选择最具有代表性的属性子集，使其保留原有数据的大部分信息，即所选择的属性子集可以像原来的全部属性一样用来正确区分数据集中的每个数据对象。一般情况下，属性选取算法主要涉及两大步：去掉与目标属性不相关的属性，删除冗余属性。

[目前DMP中仅有用户的消费属性，没有用户的本体属性，环境等,所以本部分属性选取主要目的是找出属性的重要程度

消费偏好（列表），兴趣爱好（列表），渠道，广告尺寸，消费等级，地域，创意类型，媒体类型（列表）]

[注意：不同的交换网络，价格水平会有所不同，故其作为一个属性参与分类]

对称不确定性是度量属性相关性的基本方法，定义如下：

 （3.1）

其中为信息增益(Information Gain)，计算方法为



为熵(Entropy)，为条件熵，计算方法如下：





上式中表示属性取第个值的概率，表示条件概率，即属性取值为时属性取值为的概率。

本文采用FCBF[3] (Fast Correlation-Based Filter)方法选择属性子集，该方法是处理离散型数据集属性选择的经典方法之一，采用对称的不确定性来度量两两属性之间的相关性，对处理属性数目较大时非常有效。其基本原理描述如下：设数据集有条记录，且每条记录由个非目标属性和一个目标属性来刻画。如果非目标属性与目标属性之间的相关性过低，则将该属性作为不相关属性去除。如果两个非目标属性之间的相关性过大，超过了这两个属性与目标属性的相关性时（给定阈值），则认为两个属性之间存在冗余，删除与目标属性相关性较低的属性。算法流程如下：

**输入：**训练数据集，阈值

**输出：**属性子集

1. 开始
2. for todo
3. 计算每个属性与目标属性之间的对称不确定性
4. 若，将添加到
5. end for
6. 将中的属性按的降序排序
7. for todo
8. for todo
9. 计算，若，将从属性集中去除
10. 直到中的所有冗余属性被去除，
11. end for
12. 结束

**注：**目标属性的选取可以通过计算各个属性的信息增益率来获得，选取信息增益率最大的属性作为目标属性。

#### 3.3.2属性编码

按照属性的重要度，建立分层结构，确立历史数据的统计路径。利用下图的方式建立样本编码库，在该编码库的基础上进行统计。

首先按照属性的重要度选择属性并排序，每个属性类别有各自的取值，根据取值范围给每个属性分配位数，然后根据样本的实际取值填充各自占位，前后相连即可得到每个样本的编码。

图3.3 样本编码方法

举例：

选取“性别”（取值：男，女），“年龄”（取值：0-8,19-25，25-30,31-40，40以上），“婚姻状况”（取值：未婚，已婚）作为入选属性。可见“性别”和“婚姻状况”属性分一位可完成编码，“年龄”需要四位进行编码。

表3.1“性别”属性编码

|  |  |  |
| --- | --- | --- |
| 属性 | 男 | 女 |
| 第1位 | 1 | 0 |

表3.2“年龄”属性编码

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 属性 | 0-18 | 19-25 | 25-30 | 31-40 | 40以上 |
| 第1位 | 1 | 0 | 0 | 0 | 0 |
| 第2位 | 0 | 1 | 0 | 0 | 0 |
| 第3位 | 0 | 0 | 1 | 0 | 0 |
| 第4位 | 0 | 0 | 0 | 1 | 0 |

表3.3“婚姻状况”属性编码

|  |  |  |
| --- | --- | --- |
| 属性 | 未婚 | 已婚 |
| 第1位 | 1 | 0 |

对于样本：“男，19-25，未婚”可以编码成：“101001”

对于一个样本多个取值的情况，可以采取按位相加的方式。该种编码方法的优点是对于每一个类别型属性均可以得到等长的唯一编码。但是对于稀疏样本的适应性不是很好，需要进行平滑处理。

文献[1]中提到一种“\*”树的方式对历史数据和样本经行平滑处理，带来的问题就是在做匹会有多条路径与样本或完全匹配或“\*”匹配，需要提供一种打分机制来决定样本与哪条路径是最佳匹配。该方法并不能完全适用于本文的编码结构，借鉴该种方法，在每个属性的类别中添加一个“\*”字段，这样可以通过不完全匹配以实现稀疏数据的平滑。

因为增加了“\*”属性，每个属性的编码位数和编码表语言进行改变如下：

表3.4“性别”属性编码

|  |  |  |  |
| --- | --- | --- | --- |
| 属性 | 男 | 女 | \* |
| 第1位 | 1 | 0 | 0 |
| 第2位 | 0 | 1 | 0 |

表3.5“年龄”属性编码

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 属性 | 0-18 | 19-25 | 25-30 | 31-40 | 40以上 | \* |
| 第1位 | 1 | 0 | 0 | 0 | 0 | 0 |
| 第2位 | 0 | 1 | 0 | 0 | 0 | 0 |
| 第3位 | 0 | 0 | 1 | 0 | 0 | 0 |
| 第4位 | 0 | 0 | 0 | 1 | 0 | 0 |
| 第5位 | 0 | 0 | 0 | 0 | 1 | 0 |

表3.6“婚姻状况”属性编码

|  |  |  |  |
| --- | --- | --- | --- |
| 属性 | 未婚 | 已婚 | \* |
| 第1位 | 1 | 0 | 0 |
| 第2位 | 0 | 1 | 0 |

对于样本：“男，19-25，未婚”可以编码成：“100100010”

注意：各个不同的Ad Exchange（以下统称为渠道）的平均价格水平会有所不同，所以不同的渠道也需要作为一个属性参与编码。

[给竞价活动一个合理的类别划分（即分类）方法（在影响广告价格的属性基础上），在此基础上进行竞价价格预测，需要对用户的属性（广告目标）做更细化的分析]

### 离线部分

将竞价历史数据依据时间分成三个阶段，记成，其中时期的数据用于属性选取及编码。时期的数据用于构建子样本编码库，及统计各子样本的均值，方差等信息·，构建预测模型，预测子样本竞价分布情况等。 时期的数据用于验证模型的预测精度。

[统计某类属性集下的竞价历史价格]

#### 3.4.1 构建子样本编码库

利用上一节中构建的子样本编码库，遍历时期的竞价成功数据，将其映射到子样本编码库中的编码上，并在该子样本编码上记录获胜的成交价格等信息，遍历完成后，对子样本编码上的信息进行统计。

主要包括以下几步：

1. 样本拆分，如果一条数据中某个属性具有多个取值，需要根据取值个数拆分成多个子样本，如：某条数据的各属性取值如下：年龄：18-23，性别：男，兴趣爱好：旅游，运动。可以拆分成多个子样本，S1：年龄：18-23，性别：男，兴趣爱好：旅游；S2：年龄：18-23，性别：男，兴趣爱好：运动。将每一个子样本都映射到样本编码库中的子样本编码上，并在该子样本编码上记录该条数据包含的信息，如：成交价格等。
2. 子样本编码库修剪，遍历完时期的数据后，如果某些子样本编码上所含的数据较稀疏，则该子样本编码上的数据不能真实的反映出此类样本的竞价分布情况，因此需要根据设定的阈值（如：包含的竞价数据小于200等）对该类子样本编码做丢弃处理。
3. 统计各子样本编码上的竞价数目，成交价格的均值，标准差及每层属性包含的竞价数目等信息。

#### 3.4.2 子样本竞价分布预测

统计各子样本的均值和标准差等信息，假设子样本的竞价分布服从对数正态分布，其参数和如下：





[该分布假设来自文献[1],在目前的划分下是否仍满足这一规律需要数据验证]

因此，该子样本的竞价分布可以表示如下：

 （3.2）

 （3.3）

其累积分布函数为

 （3.4）其中：



#### 3.4.3 活动竞价价格预测

假设竞价活动可以拆分成个子样本，通过子样本编码库获得各子样本的竞价分布后，根据指定成功率，以及各子样本上的分布情况，计算出竞价价格。假设每个子样本服从对数正态分布，分布参数为，指定成功率，根据公式（3.4）计算出价各子样本价格如下：

 （3.5）

假设各子样本对应的权重为，则该活动的出价计算公式如下：

 （3.6）

其中，子样本权重计算方法如下：



其中表示的是第个子样本上的竞价数目，表示的是第个子样本各属性值对应的权重之积（如果属性值对应的权重不存在，默认值为1）。

[此模型预估需要数据量在亿级别的竞价成功PV作为基础]

**注：**如果拆分出的子样本在编码库中不存在，则需要计算该子样本与编码库中所有子样本的相似度，采用相似度最高的子样本的竞价分布来代表该子样本。子样本相似度计算公式如下：



其中表示属性的个数，表示子样本的第d个属性的取值，



表示d层属性的权重，计算方法如下：假设经过各层的竞价数据个数分别是，则



注：统计竞价数据个数时是不考虑\*的情况。

### 在线部分

首先根据竞价条件，进行样本拆分，对拆分出的子样本进行编码，然后查找该编码上是否存在竞价列表，如果存在，利用竞价列表进行出价，否则，需要根据3.3.3节的预测方法计算出活动的竞价价格初值BID，再根据广告活动的差异性进行调整，调整方法如下：

考虑到每个广告活动的差异性，通用的模型在个别广告活动上会有一定的偏差，需要“实时”（目前系统下可以以天为单位）统计投放成功率，并记录每个广告活动的调整系数，该调整系数初始值为0，根据历史数据和预期成功率之间的比例关系进行调整

 （3.7）

其中为第i个广告活动的出价，为调整系数，通过预期成功率与实际成功率的比率计算而来：

 （3.8）

[根据数据实际情况，该调整系数可能不是（3.8）式中的线性关系]

[\*需要在系统中在每个广告活动维度上记录调整系数,并在竞价时实时查询]



图3.4 在线部分处理流程图

### 在线竞价记录及使用

根据业务的特点，某个用户连续在某个网站的浏览行为的捕捉及及时出价对广告的效果有很大的影响，故增加实时竞价记录及使用环节，以利于更迅速的“买到”这个类型的用户。

在系统中缓存半个小时或一个小时的竞价数据表A.1.2.1，当有新竞价请求时，首先查询该份实时竞价数据，如果按照“渠道+媒体+用户”作为联合键值，查到竞价结果则：

如果上次竞价成功，按照成交价继续竞价；

如果上次竞价失败，采用上次出价的1.2倍进行出价；

如果未能查询到竞价结果则采取利用历史出价的方式出价。如果不存在历史竞价信息，则采用系统默认出价（参见A.4）。

### 竞价价格评估与优化

通过预期成功率与实际成功率进行比较，验证竞价价格的合理性。同时可以参考竞价价格和成交价格之间的差异来评估整体出价的水平是否合适。

优化方法：

1）算法优化：不断的通过数据统计修正模型

2）基于目标的优化：根据投资回报率指标优化

3）基于规则的优化：根据制定的规则进行优化

4）调整整体的出价水平，给出价也乘以一个修正参数。

## 转化率（eTR）预测

点击率指标在广告选择时有重要的应用，所以需要对本次竞价的转化率进行预测，方便后期进行广告的选择。预测方法采用文献[2]中逻辑回归方法，根据因变量的不同得出点击转化率和Action转化率，并进行预测。本节详细算法参见文献[2]。

### 离线部分

根据竞价成功流水数据构建逻辑回归模型，计算各个属性所占权重。得到权重向量。根据目标变量的不同，可以给出点击预测模型和转换预测模型。

### 在线部分

根据竞价的媒体，广告和用户信息计算该次竞价的转换率。

基于广告活动的转化率统计数据的利用。转化率的通用模型在各个具体的广告活动上的表现可能是千差万别的，需要将基于全部历史数据生成的转化率模型在广告活动维度上做适当的修正，才能更好的反应转化率的情况。目前考虑到的方法：

通用模型计算出本次竞价的，结合系统平均转化率及广告活动的平均转化率计算出本次在广告活动上的预测转化率

 （4.1）

## 用户价值估值体系

在DMP系统中，记录着用户访问系统能采集的网站的行为信息。在这些行为信息的基础上，采用数据分析与挖掘的方法得到用户属性的知识库。该知识库为用户的价值分析提供了数据基础。

每个用户自然的拥有其自身的价值，如何建立一个用户的估值体系和方法是本章研究的主要问题。

需要说明的是，目前系统可以收集到用户的浏览，点击和订单数据这给用户的价值评估提供了一个良好的基础，但是也由于系统仅仅收集到用户的浏览，点击和订单行为，而且这些行为的跨网站的数据，造成了行为的不连续性，片面性，这也为用户的整体估值造成了很大的障碍。

图5.1展示了人群估值体系的框架。

在在线广告范畴内，用户的市场价值体现在如下几个“行为”：Impression（I），Click（C），Conversion（N）。在一个用户的生命期内，各种行为的会全部或部分，一次或多次出现。如何在这种些行为的基础上给出用户的基本价值是一个富有挑战性的工作。

基础价值的公式可以表示成：

 （5.1）

其中P(I),P(C),P(N)表示带来展现价值V(I),点击价值V(C),转化价值V(N)的可能性，由用户的历史数据统计得出。

[此处需要为每次展示，点击和转化定价，在数据收集时需要考虑这个信息，简化的做法可以是：给I和C一个“通用”价格，N如果是由商品的售卖产生的可以用订单中的价格]



图5.1 人群估值体系框架

用户的价值还包含一个重要的方面：用户的行为。这里的行为包含长期的行为和近期的行为，而且往往近期的行为会对用户的下一步操作产生更强烈的影响。表现为对某种（类）商品的购买可能性。

### 基础价值（UV\_BASE）

首先根据数据特点及广告人群类目体系，选取对用户估值重要程度高的属性进行行为识别，这是整个人群估值的数据基础。

用户特征是标识一个用户价值的有效手段，用户特征主要包括两个方面：用户的人口属性(Demographic Property)及行为定向（Behavior Targeting），根据国际推销专家[海英兹·姆·戈得曼](http://wiki.mbalib.com/w/index.php?title=%E6%B5%B7%E8%8B%B1%E5%85%B9%C2%B7%E5%A7%86%C2%B7%E6%88%88%E5%BE%97%E6%9B%BC&action=edit)(Heinz M Goldmann)对人的购买欲望程度的层次评级分成A为Attention，即引起注意；I为Interest，即诱发兴趣；D为Desire，即刺激欲望；最后一个字母A为Action，即促成购买，该分级方式同样适用于网络广告领域。

表5.1 用户属性总表

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 用户属性 | 名称 | | | 重要度分析  （低1-5高） | 可分析性  （易1-5难） |
| 本体属性 | 自然属性 | 性别 | | \*\*\*\*\* | \*\* |
| 年龄层 | | \*\* | \*\*\*\*\* |
| 婚姻状况 | | \*\*\* | \*\*\*\*\* |
| 教育水平 | | \*\* | \*\*\*\*\* |
| 社会属性 | 职业身份 | | \*\*\* | \*\*\*\* |
| 行业 | | \*\* | \*\*\*\*\* |
| 个人收入 | | \*\* | \*\*\*\*\* |
| 家庭收入 | | \* | \*\*\*\*\* |
| 关键人生阶段 | | \*\*\* | \*\*\*\*\* |
| 活动相关 | 活动参与度 | | \*\*\* | \*\*\* |
| 广告形式偏好 | | \*\*\* | \*\* |
| 兴趣属性 | Attention | 兴趣类型（浏览） | | \*\*\* | \*\* |
| Interest | 有兴趣的商品（类型）（点击） | | \*\*\*\* | \*\* |
| Desire | 有购买意向的商品（类型） | | \*\*\*\* | \*\* |
| 消费属性 | Action | 消费偏好（列表） | | \*\*\*\*\* | \*\* |
| 消费偏好权重（列表） | | \*\*\*\*\* | \*\* |
| 消费次数 | | \*\* | \*（统计报表） |
| 消费频率 | | \*\* | \*（统计报表） |
| 消费总额 | | \*\*\* | \*（统计报表） |
| 平均消费水平 | | \*\*\* | \*\* |
| 线上消费习惯 | | \*\*\* | \*（统计报表） |
| 消费者等级 | | \*\*\* | \*\* |
| 消费理念 | | \*\*\*\* | \*\*\*\* |
| 有兴趣商品 | | \*\*\*\* | \*\* |
| 有意向购买商品 | | \*\*\*\* | \*\* |
| 决定购买商品 | | \*\*\*\* | \*\* |
|  | 综合 | 时间特征 | 上网时段 | \*\*\* | （统计报表） |
| 在网时长 | \*\*\* | （统计报表） |
| 浏览时段 | \*\*\* | （统计报表） |
| 兴趣时段 | \*\*\* | （统计报表） |
| 消费时段 | \*\*\* | （统计报表） |
| 访问环境 | 设备相关 | 终端设备 | | \* | \*（统计报表） |
| os | | \* | \*（统计报表） |
| 浏览器 | | \* | \*（统计报表） |
| 语言 | | \* | \*（统计报表） |
| 地域 | 地理位置 | | \*\*\*\*\* | \* |

上表仅是近期对人的属性提取优先级比较高的属性，在这些属性分析的基础上可以得到一个人的基础价值。

目前最直接反应用户估值的公式可以表示成式（5.1），即通过用户直接参与的给广告主带来价值的行为给用户做一个基础的估值。

对于没有或缺少上述三种行为的用户，采用的方法通过人群属性构建距离公式，找到有上述数据的最近“邻居”的方式给出人的基础价值。

[目前预计有订单的人与总人数比在千分之一至万分之一这个量级上，这里涉及到稀疏数据处理的问题]

### 行为价值(UV\_BT)

在互联网广告环境中，用户的行为序列分析也是评价用户的价值的一个重要方面，所以还要辅助以行为序列识别与分析手段。如：访问特定广告的次数，来访的路径识别，搜索等等。

DMP系统对人的属性进行分析，并给出对某类（或某个）商品（commodity）的喜好程度()。序列分析会增大或减少人群对某类（或某个）商品的喜好程度，所以公式修正为[长期行为和短期行为，需要DMP做相应计算]：

 （5.2）

则：

 （5.3）

[该部分计算需要准实时的支持，新客老客应该也在此处计算出]

[影响用户对某个广告活动的兴趣程度的因素包括：

用户的消费倾向，兴趣爱好，搜索历史（搜索兴趣）

广告活动产品/类别

用户是否浏览/购买了本产品或相关产品

产品是否周期性产品-购买周期

消费等级-产品消费等级/档次

]

[涉及到的分析包括：浏览/购物日志分析->兴趣爱好和消费倾向

广告产品信息收集：类别，档次，价格等

搜索日志分析

本体属性识别\*]

[后期会加入行为系列分析]

[最终得到广告活动商品与用户的匹配程度（0,1）区间，对用户的基本价值做相应的修正]

### 竞价价值(UV\_BID)

在综合价值的基础上，根据喜好程度，商品与媒体的匹配程度等可以给出对于本次竞价，用户的价值。所以在的基础上应考虑商品与媒体的匹配程度，则公式（5.3）转化为：

 （5.4）

用户的基础估值由式（5.1）计算得出，乘以购买某类（或某个）产品的概率，即可得到该用户对本次广告的估值（）。

 （5.5）

考虑到只有竞价成功才会有花费，所以此处不必考虑竞价的成功率。

[此部分需要对广告活动进行细化，至少到商品类别层次，需要离线计算针对各个广告活动，每个用户的潜在价值]

### 估值体系评估与优化

**评估方法I**：

对用户估值体系的评估目前还没有成熟的方法进行验证，可行的方法就是通过估值（UV\_BID）与用户市场价格（BID）在人群和广告两个维度上的统计曲线比较进行，曲线应该相关且不断逼近。[这个评估方法有问题，这样系统的利润会不断的收缩]

**评估方法II**：

对用户竞价估值进行统计，并与实际的展示，点击，转化行为统计结果比较。两者应该逐渐接近，达到一定的误差范围。

如果误差超过一定的范围，有两种优化手段：

1. 调整影响估值的属性的权重，此方法需要对影响估值的属性进行重新分析。
2. 调整整体的估值水平，表现为乘以一个调整系数。

[效果数据的收集，形成从展示到转化的收集路径]

[随着RTB在广告市场的占有率的逐步增加，竞价价格也同样会不断的接近人的估值，同样会造成利润的不断减少]

## 竞价

### 指数计算

为满足平台收益和广告主的利益要求，采用竞价指数的方式对广告竞价列表进行选择。在考虑竞价指数时主要参考如下的维度：

1. 预算情况
2. 用户竞价估值
3. 广告主收费
4. 转化率
5. 竞价成功率
6. 其它（调整系数）

其中预算情况采用预算指数（BugetI）来表示，其计算公式：

 （6.1）

另由于竞价指数如此计算区间较大，为平衡各个参数的影响，对该参数限制如下：

如果该参数小于0.1则取0.1，如果该参数大于10.则取10.

因有些广告活动不会设置每天预算，故对于不设置每天预算的广告活动目前采取设定消耗指数默认值的方式，目前设定为7.5（可配置）。

则竞价指数可以表示成：

 （6.2）

每个活动的广告主出价值设为AdPrice等于UV\_BID时，计算每个活动选取不同平台出价BID时对应的竞价指数CPI，其中BID对应的成功率必须大于渠道设置（A.2.2）中对应的渠道成功比例。

另外，（AdPrice-BID）/AdPrice表示平台收益率，其数值应大于平台收益预期，初值设为20%。CPI必须大于竞价指数默认最小值，否则放弃本次出价。

[关于实时数据使用及其它策略]（此处涉及到竞价方式字段）

1 广告筛选时，标志为实时筛选成功；

2 在广告选择计算竞价指数时，增加分值。

采用的方式如下：

1. 调整预期成功率

各个渠道有设置的默认预期成功率eSuccRatio。根据各种匹配策略调整预期成功率，公式表示如下：

eSuccRatio += （1-eSuccRatio）\*α

|  |  |  |
| --- | --- | --- |
| 匹配策略 | | α |
| 实时数据 | 搜索 | 0.5 |
| 注册 | 0.7 |
| 订单 | -0.5 |
| 回头客 | | 0.2 |
| 新客 | | 0.6 |
| 商品关联 | | 0.4 |

1. 调整系数

因实施竞价列表的参与，是的价格的出价不全是由预期成功率来决定，所以可以通过调整系数来修订出价的优先级，调整策略如下：

|  |  |  |
| --- | --- | --- |
| 匹配策略 | | 调整系数 |
| 实时数据 | 搜索 | 1.5 |
| 注册 | 1.7 |
| 订单 | 0.5 |
| 回头客 | | 1.5 |
| 新客 | | 1.6 |
| 商品关联 | | 1.4 |

基于目前的情况，采用2）方法进行策略相关调整。

### 竞价方法

对上述每只广告活动本次竞价指数排序，计算竞价指数最高的活动广告出价AdPrice，使得该活动的竞价指数刚好等于第二名的竞价指数，计算得结果加一分钱，作为实际AdPrice收取广告主。

示例：

compI1=UV\_BID1 \* (AdPrice1-BID1) \* ctr1 \* bidsucRation1 \* BugetI1\*调整系数=18

compI2=UV\_BID2 \* (AdPrice2-BID2) \* ctr2 \* bidsucRation2 \* BugetI2\*调整系数=19

UV\_BID2\* (x- BID2) \* ctr2 \* bidsucRation2 \* BugetI2=18

x=x+0.01

当然此时（x-BID2）/x仍需大于平台收益预期，否则采用满足平台收益预期的x作为收取广告主价格。

## 竞价广告的选择策略[此章由第六章替代]

在计算出的待出价广告列表中，制定一个打分方案对其打分（RANK），并根据打分的分值对广告进行最终的选择。这里涉及到的因素在计算RANK是表现为加分或减分，参数需要根据实际情况进行适当的修正。RANK的初始值（）设定为一个正整数（如=1000）。

 (6.1)

### 策略选择优化

根据回头客，新客，商品关联等策略优化：

 (6.2)

其中：回头客分值，：新客分值，：商品关联分值。

根据盈利预期优化：

 (6.3)

其中eTR为此次展示的转化期望转化率（由第4章方法给出），是广告带来的价值（该价值需要在系统中设定），为单位价值到RANK的转化系数。

由于回头客在系统中的特殊地位结合产品的发展策略，对回头客采取特殊的算法，其打分方法会针对策略做特殊优化。

在出价时，仍然采用前文的分类方法，分类属性选取广告的目标属性；

估值时：因缺少人的信息，估值采用出价乘以系数方式给出

打分：有回头客时，回头客优先。

### 广告池优化

广告池内广告展现分布优化：

 (6.4)

其中为该时段期望展示的次数，为该时段实际展示的次数，为展示在打分中占的分值。

### 用户估值投票

如果用户的竞价价值（UV\_BID）低于本次竞价的成本，则放弃该只广告的本次竞价机会。在系统运行前期，由于用户估值的准确度未经严格的验证，此指标与竞价价格之间的差值作为广告竞价打分的参数，而不作为一票否决的机制使用。

 (6.5)

其中为用户估值在打分中所占分值。

### 总体选择优化方案

经过（6.1）~（6.5）打分计算后，取RANK最高者出价（当最高RANK低于某一阈值时，放弃此次竞价）。这种RANK方法优点是直观明了，且可以在后续增加影响RANK的分值项，缺点是很多参数需要设置默认值，这些默认值需要在业务分析和数据分析的基础上给出并不断调整。

[可行的实施方法是：根据重要性一步一步添加到计算公式中，不断验证参数正确性]

实际应用中，不同阶段采用不同的权重系数来计算竞价指数，竞价指数高者获得出价的机会。

### 广告集合竞价

当系统中广告数量增加到一定数目时，应加入本系统内广告集合竞价的手段，调整每只广告的价格，以期达到更好的投放效果和投放收益。

在内部广告活动较多时，可以通过竞价指数（CPI）完成内部竞价，调整广告的出价以获得出价机会。

## 数据输入输出及系统需求

### 数据输入

#### DMP数据需求

用户属性特征：

详见A.1。

用户行为特征：

1） 用户的搜索行为数据

2） 用户的特定网站访问行为数据

3） 消费品位等\*

后期可以考虑增加实时属性分析部分，做到属性更新的快速性。

#### DSP数据需求

1. 广告主信息
2. 广告活动信息：

目标人群，投放类型（AdServiceType），最高出价（AdMaxBid），（商品或活动）价值(AdValue)，[档次]，商品类别，广告的创意类型(AdCreativeType)，尺寸(AdSize)，预期成功率（eSuccessRatio）的设置与调整接口。

3）渠道设置表。

详细字段参见A.2。

#### 竞价数据需求

1. 离线部分

参见A.3.1。

1. 在线部分

用户属性，媒体属性，广告活动属性，是否回头客（回头客属性），是否商品关联。

参见A.1，A.3.1，A.2.1，A.2.2。

#### 系统需求

作为一个实时竞价系统，其获取的数据的实时性在很大程度上影响到出价的反应速度，另外由于算法的复杂性，对算法的完善和更新也是该系统必须面临的问题。故在对系统的要求除上述数据属性的要求外，还包含下面两方面的需求：

1. 数据采集，分析和利用的实时性。
2. 在系统级别实现对算法上线，修改等操作可以在线流量切换的功能。
3. 为满足数据积累是产生的大数据量问题，系统应具备增量数据处理手段。特别对于DMP系统中的各个属性，应能给出不同时间段的分析结果。如：消费偏好，至少应可以给出长期偏好和近期偏好（一月，一周，一天，一小时等）。

### 数据输出

#### 竞价输出

出价的广告活动，价格，渠道，策略。

#### 算法效果验证输出

1）竞价过程中输出

用户ID，用户基础价值（UV\_BASE），用户竞价估值（UV\_BID）（列表），竞价价格（BID）（列表），竞价打分（列表），出价活动，出价价格（BID），预期转化率（eTR），预期成功率（eSucRatio）。记录在表A.3.1中。

2）业务系统支持

出价活动，出价价格（BID），竞价成功率（aSucRatio），用户竞价价值（UV\_BID）（收广告主价格），预期转化率,出价策略。记录在表A.3.1中。

1. 竞价结果统计分析报表

详见表A.3.2。

评价方法：预期值应趋近实际值。

4）竞价算法效率分析数据

每次竞价计算的周期，耗时最长的计算步骤等，此部分应打不同级别的log，在算法测试期打开，生产状态可以逐步关闭。从竞价请求到给出出价活动和价格的计算过程时间消耗应小于20ms。

#### 活动目标修正统计

由于活动的目标人群是人工指定的，这种设定有一定的任意性，并不一定是十分合理的。所以为不断提高活动的投放效果，需要对活动展示，点击和转化的人群进行统计。该统计数据主要用于评价广告活动投放效果并给优化人员一个指导性的意见用以更改活动目标人群。

## 参考文献

[1] Ying Cui, Ruofei Zhang, etc, Bid Landscape Forecasting in Online Ad Exchange marketplace, KDD’11, August 21-24, 2011, San Diego, California, USA

[2]王曼，苏文魁，RTB实时竞价算法，<http://172.16.10.50:8092/svn/RandD/XinMode/Doc/rtb/RTB实时竞价算法.docx>

[3] L. Yu and H. Liu. Feature selection for high-dimensional data: A fast correlation-based filter solution. In Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003), pages 856-863, 2003.

[4] G. J. McLachlan and D. Peel. Finite Mixture Models. New York: Wiley, 2000.

[5] DMP产品需求说明书\_v1.95

## 附录A 数据输入输出详细字段

### A.1 DMP数据需求

#### A.1.1用户属性数据

##### A.1.1.1长期属性表

|  |  |
| --- | --- |
| **序号** | **字段名称** |
| 1 | 用户标识 |
| 2 | 平台ID |
| 3 | 兴趣爱好 |
| 4 | 兴趣爱好权重 |
| 5 | 消费偏好 |
| 6 | 消费偏好权重 |
| 7 | 消费总次数 |
| 8 | 消费频率 |
| 9 | 消费总额 |
| 10 | 平均消费水平 |
| 11 | 线上消费习惯 |
| 12 | 消费者等级 |
| 13 | 地理位置 |
| 14 | 终端设备 |
| 15 | OS |
| 16 | 浏览器 |
| 17 | 有兴趣商品 |
| 18 | 有兴趣商品权重 |
| 19 | 有意向购买商品 |
| 20 | 有意向购买商品权重 |
| 21 | 决定购买商品 |
| 22 | 决定购买商品权重 |
| 23 | 用户基础估值 |

提供周期，每天更新一次，全量取。

##### A.1.1.2 实时属性表

|  |  |  |
| --- | --- | --- |
| **序号** | **字段名称** | **来源** |
| 1 | 用户标识 |  |
| 2 | 平台ID |  |
| 3 | 消费偏好 | 实时订单及注册 |
| 4 | 消费总额 | 实时订单及注册 |
| 5 | 有兴趣商品 | 实时搜索 |

实时更新，支持实时查询。

#### A.1.2竞价数据

##### A.1.2.1竞价流水数据

|  |  |  |  |
| --- | --- | --- | --- |
| **序号** | **字段名称** | **类型** | **源** |
| 1 | Request ID |  | RTB |
| 2 | 广告ID |  | RTB |
| 3 | 渠道ID |  | RTB |
| 4 | 媒体类型 | 列表 | RTB |
| 5 | 权重 | 列表 | RTB |
| 6 | 广告位宽 | 整型 | RTB |
| 7 | 广告位高 | 整型 | RTB |
| 8 | 用户ID |  | RTB |
| 9 | 竞价方式 |  | RTB |
| 10 | 竞价价格 |  | RTB |
| 11 | 用户价值 |  | RTB |
| 12 | 预测转化率 |  | RTB |
| 13 | 预测成功率 |  | RTB |
| 14 | 是否成功 |  | DSP |
| 15 | 成交价格 |  | DSP |
| 16 | 是否点击 |  | DSP |
| 17 | 时间 |  | RTB |

该表由DMP关联来源于RTB和DSP的数据，时间周期，按需求提供，一般为一星期或十天为周期。（计算转化率模型）

在实时竞价中，会缓存该表一定时期。

##### A.1.2.2统计数据

**竞价历史统计（计算与分析用-广告维度-时间段内统计数据）**：

|  |  |  |
| --- | --- | --- |
| **序号** | **字段名称** | **类型** |
| 1 | 广告ID |  |
|  | 活动ID |  |
| 2 | 渠道ID |  |
| 3 | 竞价次数 |  |
| 4 | 成功率 |  |
| 5 | 平均成交价格 |  |
| 6 | 平均出价价格 |  |
| 7 | 点击率 |  |
| 8 | 最近成交价格 | 列表（最近5次） |
| 9 | 最近出价 | 列表（最近5次） |

数据应做到准实时性，RTB可以在准实时的状态下获取，可实时（或短期如半小时）查询。

(利用实时数据计算CTR，用户的估值要基于用户对活动的价值，目标设置调整（DSP），实施的性能（参数选取）)

**竞价历史统计（计算与分析用-媒体维度）**：

|  |  |  |
| --- | --- | --- |
| **序号** | **字段名称** | **类型** |
| 1 | 媒体ID |  |
| 2 | 渠道ID |  |
| 3 | 竞价次数 |  |
| 4 | 成功率 |  |
| 5 | 平均成交价格 |  |
| 6 | 平均出价价格 |  |
| 7 | 点击率 |  |
| 7 | 最后成交价格 |  |
| 8 | 最后出价 |  |

一天更新一次，支持实时查询。

### A.2 DSP数据需求

#### A.2.1 DSP活动表

|  |  |  |
| --- | --- | --- |
| **序号** | **字段名称** | **备注** |
| 1 | 活动ID |  |
| 2 | 活动有效性 |  |
| 3 | 活动名称 |  |
| 4 | 投放类型 | CPC或CPM |
| 5 | 活动预算 |  |
| 6 | 活动每日预算 |  |
| 7 | 投放时间 |  |
| 8 | 活动最高出价 |  |
| 9 | 媒体倾向 |  |
| 10 | 时段划分 |  |
| 11 | 地域 |  |
| 12 | 年龄 |  |
| 13 | 性别 |  |
| 14 | 婚姻 |  |
| 15 | 教育程度 |  |
| 16 | 兴趣 |  |
| 17 | 购物倾向 |  |
| 18 | 是否新客 |  |
| 19 | 是否老客 |  |
| 20 | 新客规则ID |  |
| 21 | 老客规则ID |  |
| 22 | 活动消耗 |  |
| 23 | 活动类型 |  |
| 24 | 商品类别 |  |
| 25 | 创意类型 |  |
| 26 | 广告位长度 |  |
| 27 | 广告位高度 |  |
| 28 | 活动优先级 |  |
| 29 | 广告主名称 |  |
| 30 | 广告主类型 |  |

实时更新，用于广告筛选，转化率预测，竞价等。

#### A.2.2渠道设置表

|  |  |  |
| --- | --- | --- |
| **序号** | **字段名称** | **类型** |
| 1 | 渠道ID |  |
| 2 | 渠道类型 |  |
| 3 | 渠道权重 |  |
| 4 | 竞价成功率要求 |  |

### A.3竞价数据

#### A.3.1 流水数据

|  |  |  |
| --- | --- | --- |
| **序号** | **字段名称** | **类型** |
| 1 | 广告ID |  |
| 2 | 渠道ID |  |
| 3 | 媒体类型 | 列表 |
| 4 | 权重 | 列表 |
| 5 | 广告位宽 | 整型 |
| 6 | 广告位高 | 整型 |
| 7 | 用户ID |  |
| 8 | 竞价方式 |  |
| 9 | 竞价价格 |  |
| 10 | 竞价估值 |  |
| 11 | 预测转化率 |  |
| 12 | 预测成功率 |  |
| 17 | 时间 |  |

该表每次竞价给出，实时出价及历史统计时使用。

#### A.3.2竞价分析数据表

|  |  |  |
| --- | --- | --- |
| **序号** | **字段名称** | **类型** |
| 1 | 广告ID |  |
| 2 | 渠道ID |  |
| 3 | 竞价次数 |  |
| 4 | 平均预期成功率 | DMP |
| 5 | 实际成功率 | DMP |
| 6 | 预期平均点击率 | DMP |
| 7 | 实际点击率 | DMP |
| 8 | 出价平均价格 | DMP |
| 10 | 实际平均成交价格 | DMP |
| 11 | 平均用户价值 | DMP |
| 9 | 预算消耗 | DSP |
| 10 | 实际消耗 | DSP |

该表以天为周期，按广告维度统计给出。

### A.4其它

#### A.4.1媒体维度默认值表

|  |  |  |
| --- | --- | --- |
| **序号** | **字段名称** | **类型** |
| 1 | 媒体ID |  |
| 2 | 平均出价 |  |
| 3 | 平均点击率 |  |

#### A.4.2广告维度默认值表

|  |  |  |
| --- | --- | --- |
| **序号** | **字段名称** | **类型** |
| 1 | 广告ID |  |
| 2 | 平均出价 |  |
| 3 | 平均点击率 |  |

上述两份表由产品根据市场情况给出，作为默认值和初始值使用，更新周期：长期。

## 附录B 算法调整

### B.1 关于用户估值的调整

用户的基础估值原来的每次展现和点击价格定义偏高，目前将用户的基础估值调整如下：

首先定义RTB平台的每次展现价值为：

（可配置参数）



式中（目前DMP中的用户基础每次展现价值）。如果UV\_BID<则UV\_BID=;如果UV\_BID>广告主最高限价，则UV\_BID=广告主最高限价。

### B.2 CPI指数计算

计算CPI时，由于UV\_BID、BID、AdPrice表现在单次展现上数值较小，乘以1000后参与计算；

计算CPI时，eTR采用的方式带入公式，目前A=e（2.71828），此参数为可配置参数。

公式调整为：

 如果RTB的实时数据记录中有该用户的实时竞价信息，特别是当上次竞价不成功时，会造成UV\_BID=BID的情况，使得CPI的值为0，且预期成功率为1.0。此时CPI的计算公式退化为：



### B.3 预期成功率设置

目前预期成功率设置为0.5(可配置)。

### B.4 默认点击率设置

默认点击率由原来的0.001更改为0.0005。

### B.5 关于是否首屏的参数使用

有数据表明在首屏显示的广告的平均点击率会比在第二屏的点击率高20%左右，所以竞价请求中是否首屏的参数对竞价的效果会有很大的影响。根据Google实时竞价协议(realtime-bidding-proto.txt)在竞价请求(message BidRequest)中的AdSlot字段：

optional SlotVisibility slot\_visibility = 12 [default = NO\_DETECTION];

以枚举型数据表示出广告位是否在第一屏可见，数值如下：

enum SlotVisibility {

NO\_DETECTION = 0;//未知

ABOVE\_THE\_FOLD = 1;//首屏显示

BELOW\_THE\_FOLD = 2;//首屏不显示

}

对该字段的利用方式体现在竞价价格和预期转化率两个方面。

对竞价价格BID（通过历史数据得出的情况）调整如下：

如果 slot\_visibility = 0则不作调整

如果 slot\_visibility = 1 则计算出竞价价格乘以1.1(可配置)

如果 slot\_visibility = 2 则计算出竞价价格除以1.1(可配置)

对预测转换率(eTR)调整如下：

如果 slot\_visibility = 0则不作调整

如果 slot\_visibility = 1 则计算预测转化率(eTR)乘以1.1(可配置)

如果 slot\_visibility = 2 则计算预测转化率(eTR)除以1.1(可配置)

为完成相关更改,RTB和DMP应做的相应改动：

1. RTB系统在解析竞价请求时，需要解析出slot\_visibility字段，并记入竞价流水，需注意与DMP关于竞价流水的相关接口应做相应改动（增加是否首屏的字段）。

2. RTB竞价算法计算时应按上述要求进行相应更改。

### B.6 点击率预测抽样

目前竞价成功的流水数据量过大，发生点击的数据量较小（万分之五左右），因此需要对未发生点击的数据进行抽样，假设需要抽取的未发生点击的数据量为10%（根据数据量的变化，有可能改变），抽样方法如下：

1. 判断该条数据是否发生点击，如果发生点击，抽取该条数据，否则，转2
2. 产生一个[1,100]均匀分布的随机数n，如果n < 11，选取该条记录

### B．7 点击率预测修正

考虑到实际点击率与算法模型计算出的点击率之间的关系，为使得预期点击率在可控制的范围内，对预期点击率进行修正。

 （2.4）

其中：为平均点击率，应该是可配置参数，默认为0.0005。