**全渠道导粉CRM**

**架构设计**

**陈强**

**2015-03-17**

**目录**

[1 总体架构 3](#_Toc414394255)

[1.1. 系统概要 3](#_Toc414394256)

[1.2. 设计原则 3](#_Toc414394257)

[1.3. 总体架构 3](#_Toc414394258)

[2 数据导入 3](#_Toc414394259)

[2.1. 数据类型 3](#_Toc414394261)

[2.2. 导入方式 3](#_Toc414394262)

[2.2.1. API 3](#_Toc414394263)

[2.2.2. 数据文件 3](#_Toc414394264)

[2.2.3. 爬取 4](#_Toc414394265)

[2.3. 导入辅助 4](#_Toc414394266)

[3 客户统一模型 4](#_Toc414394267)

[4 客户价值模型 4](#_Toc414394268)

[4.1. 模型设计原则 4](#_Toc414394271)

[4.2. 维度 5](#_Toc414394272)

[4.3. 算法 5](#_Toc414394273)

[5 客户跟进 5](#_Toc414394274)

[6 同人模型 5](#_Toc414394275)

[6.1. 数据 5](#_Toc414394282)

[6.2. 计算模型 5](#_Toc414394283)

[6.2.1. 模型的输入输出 5](#_Toc414394284)

[6.2.2. 计算项 6](#_Toc414394294)

[6.2.3. 公式 6](#_Toc414394295)

[6.3. 辅助技术 6](#_Toc414394296)

[6.3.1. 地址分析 6](#_Toc414394297)

[6.3.2. 注册号字符相近 6](#_Toc414394298)

[7 渠道 6](#_Toc414394299)

[8 系统设计 6](#_Toc414394300)

[8.1. 微商城 6](#_Toc414394309)

[8.2. CRM后台 7](#_Toc414394310)

[8.3. 手机端APP 7](#_Toc414394311)

[8.4. 数据分析运算系统 7](#_Toc414394312)

[8.4.1. 数据运算能力 7](#_Toc414394313)

[8.4.2. 数据运算任务 7](#_Toc414394314)

[8.4.3. 数据架构 7](#_Toc414394315)

[9. 其它 7](#_Toc414394316)

# 总体架构

## 系统概述

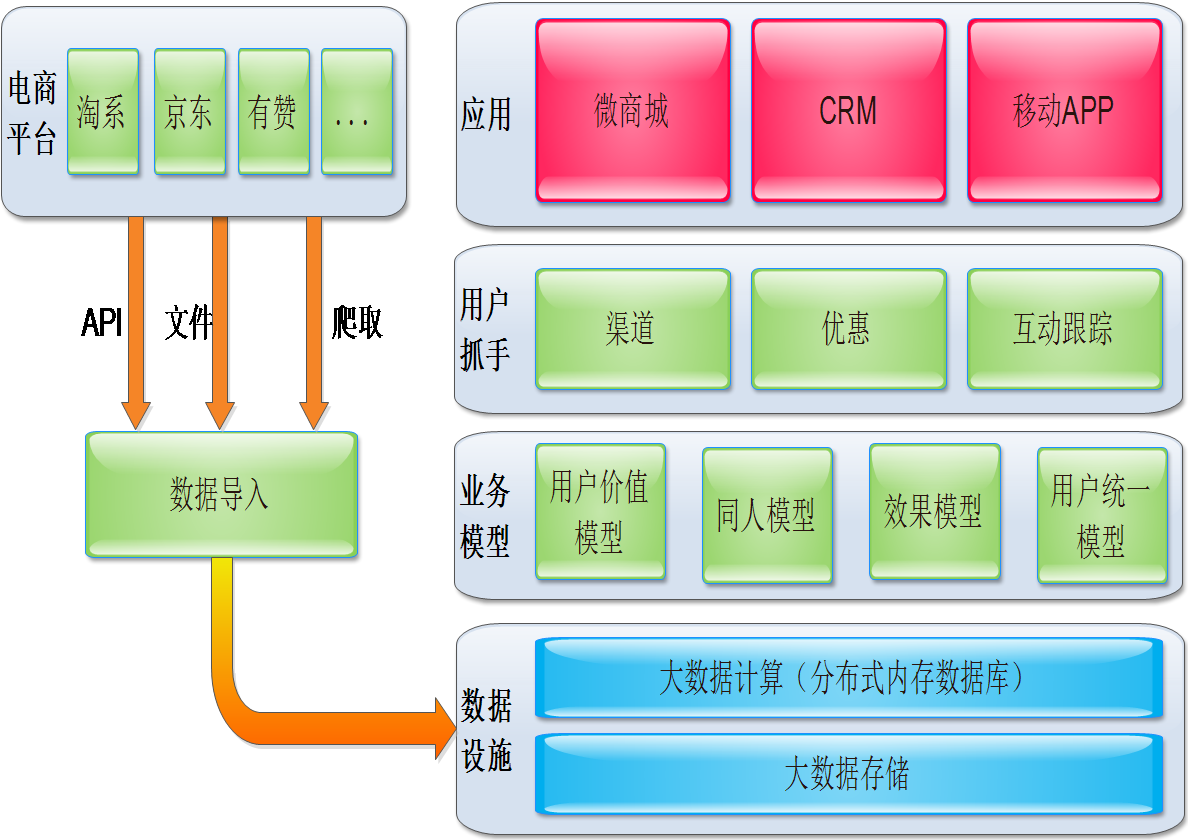
对于众多电商商家来说，拥有多个跨平台的业务，但跨平台的营销，基本是不可能，原因一是平台的天然封闭性，只要是电商平台，天然不会允许数据流失，用户流失；二是商家的能力有限，没有数据的经营能力，或者数据意识淡薄。这就导致了众多电商商家苦苦受制于平台，离开平台，平台规则变化，生意大受影响。大数据时代，数据怎么能受制于人呢？

该系统的设计研发，就是为了解决商家的数据痛苦，让商家不再受制于平台，以数据为核心展开经营。

## 设计原则

1. 快速，数据高速处理
2. 准确，数据计算误差小
3. 大，面向大商家的大数据，处理大商家的用户数据
4. 杂，横跨主要电商平台（天猫，天京，苏宁等），微信，有赞商城等
5. 闭环，数据形成闭环

## 总体架构



# 数据导入



## 数据类型

可以包含的数据有：交易，商品，会员，浏览，收藏（含购物车），优惠

暂时只考虑交易数据

## 导入方式

### API

淘宝天猫，1688，京东，唯品会，一号店，苏宁，各种ERP，速卖通，亚马逊，ebay

产品第一个版本只包含：淘宝天猫、京东

### 数据文件

淘宝天猫，1688，京东，唯品会，一号店，苏宁，各种ERP，速卖通，亚马逊，ebay

产品第一个版本只包含：淘宝天猫、京东

### 爬取

页面爬取或者浏览器插件

## 导入辅助

宝贝标题做分词

# 客户统一模型

|  |  |  |  |
| --- | --- | --- | --- |
| **Id** | **列名** | **作用** | **备注** |
| 1 | sName | 本人姓名 |  |
| 2 | sMobile | 本人手机 |  |
| 3 | sAddress | 本人地址 |  |
| 4 | sSex | 本人性别 |  |
| 5 | sRegistration | 本人注册号 |  |
| 6 | sWeixin | 本人微信 |  |
| 7 | sQq | 本人QQ |  |
| 8 | sAlipay | 本人支付宝 |  |
| 9 | sFan | 粉丝ID |  |
| 10 | rName | 收货人姓名 |  |
| 11 | rMobile | 收货人手机 |  |
| 12 | rAddress | 收货人地址 |  |
| 13 | rSex | 收货人性别 |  |
|  |  |  |  |
|  |  |  |  |

# 客户价值模型



## 模型设计原则

全渠道客户价值模型（CVM，customer value model），收集全渠道数据，基于已有的交易及其它数据，以客户未来可以给公司带来的价值为基准，计算客户价值。理想的计算结果是，未来实际的客户价值等于计算出来的预判的客户价值。模型需要依据不同的行业，不同的客户特性做调整。同时，需要根据数据做出校正，可以以近半年的数据为未来数据，以半年之前的数据为计算数据，来调校模型。

## 维度

全局维度：最近一次购买时间，购买次数

订单维度：购买金额，购买品类，渠道，退换货，评价，分享

## 算法

客户未来价值 = ∑（全局维度 \* 全局系数） + ∑（订单维度 \* 订单系数）

系数，可由训练算法来优化

# 客户跟进

# 同人模型

由于数据从不同的渠道获取，每个渠道取到的数据不一样，如何用这些不同维度的数据，来判断是否是同一人，在业务中有着广泛的应用。所以需要建立一个用不同的关于人的数据，来判断是否同一个人的数据计算模型。这个模型的设计需要遵循几个原则：1.对不同维度数据的极大兼容性与扩展性，可以任意添加新的数据维度；2.数据越多，准确性越大；3.计算结果量化，以直观的数据来判断是否同人。



## 数据

天猫数据：本人注册号，本人支付宝，收货人姓名，收货人地址，收货人固话，收货人手机

京东数据：本人注册号，收货人姓名，收货人地址，收货人手机

有赞数据：微信粉丝ID，本人注册号，收货人城市，收货人区域，收货人姓名，收货人省份，收货人地址，收货人邮编，收货人手机

全局变量：代他人购物比例，注册号字符相近，地址相近

微信数据：openid, unionid, 用户昵称，性别，城市，国家，省份，订阅时间

## 计算模型

### 模型的输入输出

输入：两条交易记录

输出：两条交易记录中的买家为同一人的概率



### 计算项

常用的计算项包括以下：注册号相等，手机号相等，同城同区同路同地址，邮编相等，姓名相等，固话相等

其中每一个计算项，都有着不同的处理逻辑，如注册号相等，即使是完全相同，也不会返回100%的概率，在不是完全相同的情况下，还要做字符相近分析。地址计算项的分析，需要做语义分析，做省市区的精确匹配，然后再做地址比较。

在做这些计算项处理时，需要一些全局变量，包括代购比例，字符相近占比等。

### 公式

假设有N个计算项，计算出的独立概率分别为 X1, X2 … Xn

再假设这n个计算项是独立概率事件，那么每一个计算项非同人的独立概率是1-Xi，于是就有了下面的同人模型：

**同人概率 = 1 – (1 - X1) \* (1 – X2) \* …. \* (1 – Xn)**

## 辅助技术

### 地址分析

TODO:

### 注册号字符相近

有三种算法：余弦定理；编辑距离；最长公共子串。具体采用哪种算法，需要依据实际的效果来定。

***余弦定理***：假如字符串s1与s2，比较两个字符串的相似度，sim(s1,s2)，假设s1,s2中含有n个不同的字符，其分别为c1,c2,...cn，判断字符串的相似度转换为两个字符串对应的向量v1,v2之间夹角大小的判断，余弦值越大其向量之间的夹角越小，s1与S2的相似度越大。

在向量空间模型中，文本泛指各种机器可读的记录。用D（Document）表示，特征项（Term，用t表示）是指出现在文档D中且能够代表该文档内容的基本语言单位，主要是由词或者短语构成，文本可以用特征项集表示为D(T1，T2，…，Tn)，其中Tk是特征项，1<=k<=N。例如一篇文档中有a、b、c、d四个特征项，那么这篇文档就可以表示为D(a，b，c，d)。对含有n个特征项的文本而言，通常会给每个特征项赋予一定的权重表示其重要程度。即D＝D(T1，W1；T2，W2；…，Tn，Wn)，简记为D＝D(W1，W2，…，Wn)，我们把它叫做文本D的向量表示。其中Wk是Tk的权重，1<=k<=N。在上面那个例子中，假设a、b、c、d的权重分别为30，20，20，10，那么该文本的向量表示为D(30，20，20，10)。

***编辑距离***：编辑距离就是用来计算从原串（s）转换到目标串(t)所需要的最少的插入，删除和替换。

***最长公共子串***：求两个字符串最长公共子串的问题。解法就是用一个矩阵来记录两个字符。串中所有位置的两个字符之间的匹配情况，若是匹配则为1，否则为0。然后求出对角线最长的1序列，其对应的位置就是最长匹配子串的位置。

## 使用场景

1. 用户在微信上绑定手机时，与其它交易平台的用户对应

# 渠道

短信，邮件，电话

# 系统设计



## 微商城

有赞

## CRM后台

导数据，查看数据，数据分析，导粉，效果分析

## 手机端APP

查看导粉数据，分享朋友圈

## 数据分析运算系统

### 数据运算能力

对于单个商家而言，他的数据计算任务，需要快速的完成，不接受10秒级别以上的等待。

以一个年GMV一亿元的商家为例，假设订单均价是100元，那么有100万笔订单，一笔订单保守按照1K数据保存，那么需要的数据容量是1G。

OK，数据量目标定为10G，RT定为10s以内。计算的要求，是需要对这些数据做任意的检索，包括语义分析等。

### 数据运算任务

数据检索

同人模型计算

客户价值计算

### 数据架构

单机内存数据库，把数据全部放在内存中

全文搜索引擎，数据可放在内存或者SSD中

TODO:

# 数据库设计

原始交易表（淘宝天猫，京东，有赞，每个数据源一张表）

统一用户表（同人模型未处理的用户表）

同人用户表（同人模型处理后的用户表），生成同人主键

同人交易表（同人模型处理后的交易表）

同人互动表（与客户的交互，包括微信，短信，电话，游戏，红包等，不包括销售客服）

公众号关注表

红包发放表

积分表

# 其它

发短信给用户，根据城市地址来关联是谁，或者在用户关注公众号，回复特定信息来判定人。或者也可以根据

在微商城上生成的订单，也可以与已有交易数据关联，来确定同人

原始交易数据存储 -> 同人模型处理（在初步阶段可忽略） -> 客户模型 -> 微商城交易数据 -> 粉丝互动数据

销售、粉丝、社群，从这三个维度来计算用户价值