**推荐系统用户画像设计文档**

版本： 0.1.0

拟制: 葛立青

**修改记录**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **文件编号** | **版本号** | **拟制人/修改人** | **拟制/修改日期** | **更改理由** | **主要更改内容** |
|  | 0.1.0 | 葛立青 | 2016-11-25 |  | 起草 |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
| 注1：每次更改归档文档时必须填写该表 注2：一次归档时，“更改理由”和“主要更改内容”填无 | | | | | |

**目录**

[1. TODO 4](#_Toc421712949)

[1.1 TODO 4](#_Toc421712950)

[1.2 TODO 4](#_Toc421712951)

[1.3 TODO 4](#_Toc421712952)

[2. TODO 5](#_Toc421712953)

[2.1 TODO 5](#_Toc421712954)

[2.2 TODO 5](#_Toc421712955)

[2.3 TODO 5](#_Toc421712956)

# 概述

## 应用项目

数据统计平台项目

## 开发背景

### 为什么需要构建用户画像

* 精准营销服务；分析产品潜在用户，针对特定群体利用短信邮件等方式进行营销；
* 数据挖掘，构建智能推荐系统；利用关联规则计算，喜欢红酒的人通常喜欢什么运动品牌，利用聚类算法分析，喜欢红酒的人年龄段分布情况；

……等等

## 术语

* 特征（feature）1

//TODO

* 内容（item）2

//TODO

# 设计

## 设计目的

本设计旨在构建努比亚用户标签库和用户画像（User Profile）。根据用户人口社会属性、业务应用使用行为习惯等信息而抽象出的一个标签化的用户模型。

## 架构

//TODO 业务层次、技术架构、部署架构

## 模块划分

推荐系统用户画像构建设计划分为：标签库管理模块、数据收集模块、行为建模模块、构建画像模块。

* 标签库管理模块：本模块用于管理标签库体系中的标签，包括人口属性标签、兴趣标签及业务内容[2]标签。
* 数据收集模块：本模块用于收集数据，包括但不限于：业务系统用户行为数据、网络流行话题等。
* 行为建模模块：包括准备数据、模型训练步骤；训练的模型，如：预测用户娱乐兴趣模型、预测用户音乐兴趣模型。
* 构建画像模块：根据各类模型，预测用户兴趣标签及权重值。

## 标签库管理模块

标签是联系用户与内容的关键要素，本模块旨在构建一个可维护更新的用户标签库。

### 用户标签库初始构建

人工导入初始用户标签库，用户标签分为三大类：人口社会属性标签、行为特征标签和兴趣爱好标签。

人口社会属性标签库：我们定为这种特征是**不会改变**的标签，注意包括心理现象（性格、气质等）

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 性别 | 年龄段 | 婚否 | 职业 | 学历 | 性格 | 资产 | 宗教 | … |
| 男 | 18以下 | 是 | 大学生 |  |  | 有车； |  |  |
| 女 | 18～22 | 否 | 研究生 |  |  | 将买房； |  |  |
|  | 22～25 |  | 工人 |  |  |  |  |  |
|  |  |  | 白领 |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |

这样的数据字典保存在关系型数据库MySQL中， 同样地，将兴趣标签也保存在MySQL数据字典里。

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 音乐 | 旅游 | 运动 | 游戏 | 娱乐 | … |
| 流行乐 | 穷游 | 足球 | RPG游戏 | 宠物 |  |
| 二胡 | 国内游 | 篮球 | 休闲益智 | 明星八挂 |  |
| 戏曲 |  | 游泳 | 体育竞速 | 搞笑 |  |
| 吉他 |  |  |  |  |  |
|  |  |  |  |  |  |

### 建立业务内容标签与用户标签关联规则

首先，业务内容标签暂时由各业务系统提供。各个业务独立，在与用户标签建立关系时。

其次，各个业务系统准备接入标签系统时，需要人工初始化导入内容标签与用户标签库关系表，其中用户标签库仅由我们提供。

最后，业务系统发布新内容并带有全新的内容标签，我们系统自动为其与用户标签库建立关联关系，具体策略如下：

### 用户标签库更新

## 数据收集模块

## 行为建模模块

## 构建画像模块