

# 用户标签应用实施方案说明书

文档记录:

摘要记录:

关键字	用户标签 权重 排序
编 号	用户标签应用实施方案说明书
关 联	

变更记录:

[illegible]

## 1. 引言

### 1.1 项目名称

用户标签应用实施方案

### 1.2 项目背景及概要

随着用户行为数据爆发式的增长，不同业务数据的应用不尽相同，用户行为数据的标签抽取也存在较大差异。聚焦个性化场景的用户个性化推荐营销，实现数据精细化运营管理，进而深入挖掘潜在的商业价值。

本文主要基于用户行为分类其中主要包括：挂号、问诊、报到、搜索、浏览、关注、加入、发回帖等，抽取行为相关的医疗标签（包括系统标签和医言堂自定义标签），通过不同的业务权重配置以及行为时间衰减系数计算，用以综合评估用户标签的得分，以便各业务的营销推荐场景的不同应用。

### 1.3 项目目标

实现用户标签数据针对不同业务项目的精准定位及个性化推荐应用。

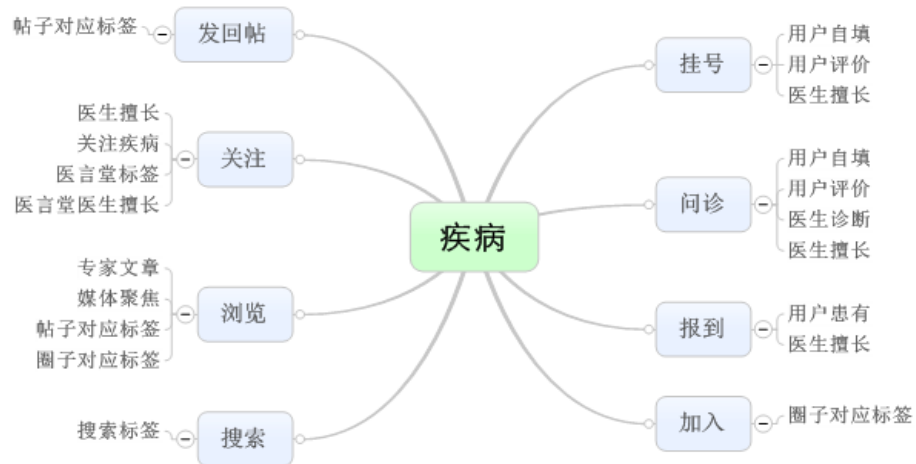
### 1.4 项目适用范围

各搜索及推荐业务场景的应用人员。

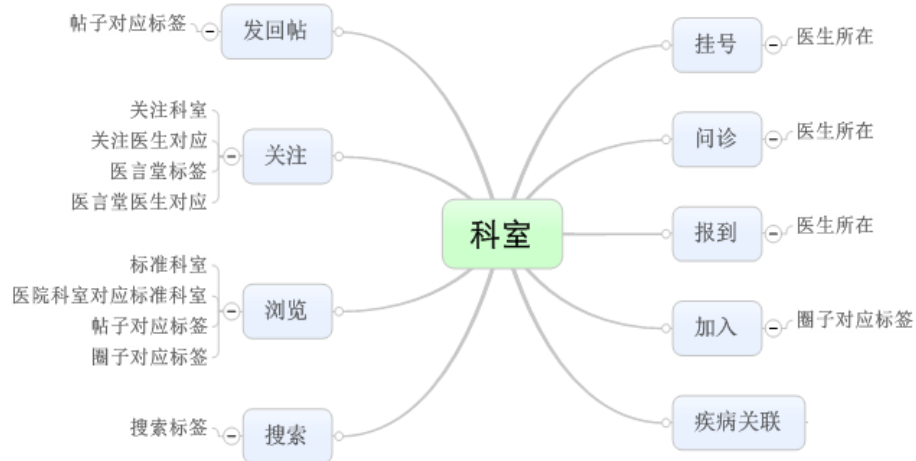
## 2. 标签开发内容

标签开发从**标签类型**分主要有疾病、药品、症状、治疗、医院、医生、科室、人体、人群、医学指标、过敏原等，从**行为**分主要有挂号、问诊、报到、关注、浏览、搜索、加入圈子、关联、发回帖九类。

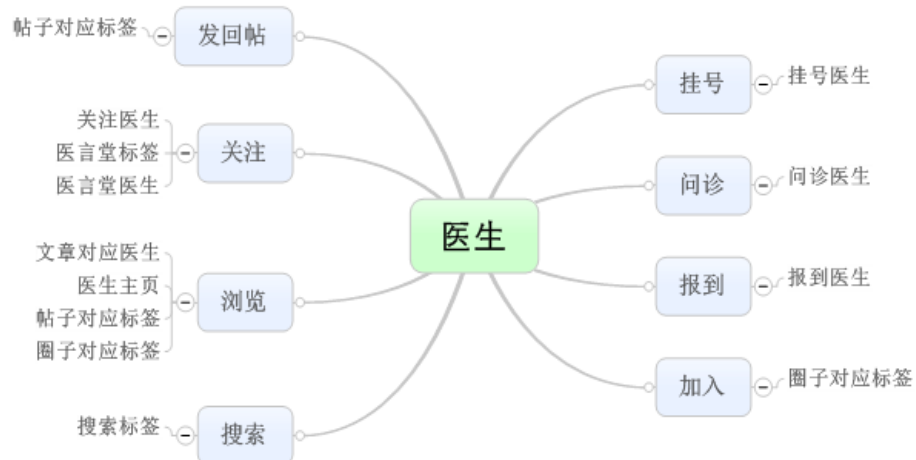
## 2.1 疾病标签



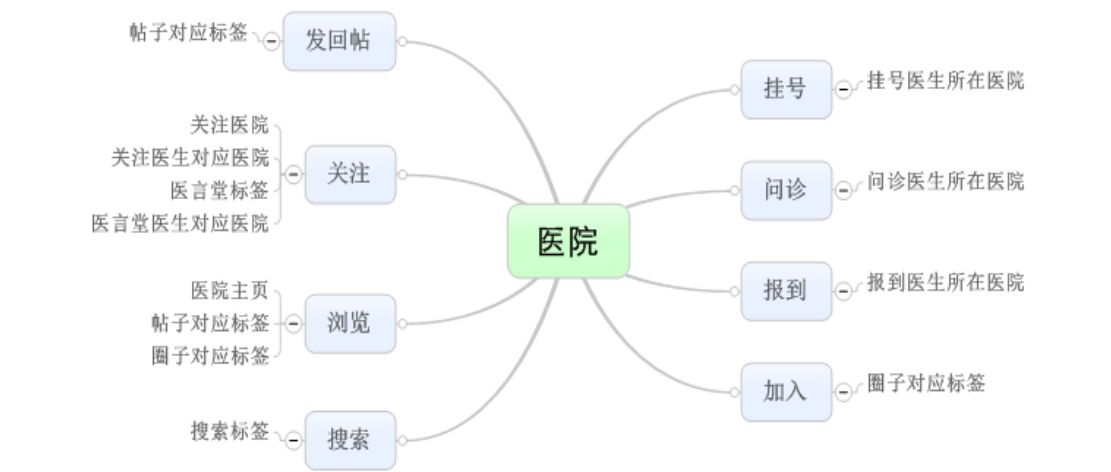
## 2.2 科室标签



## 2.3 医生标签

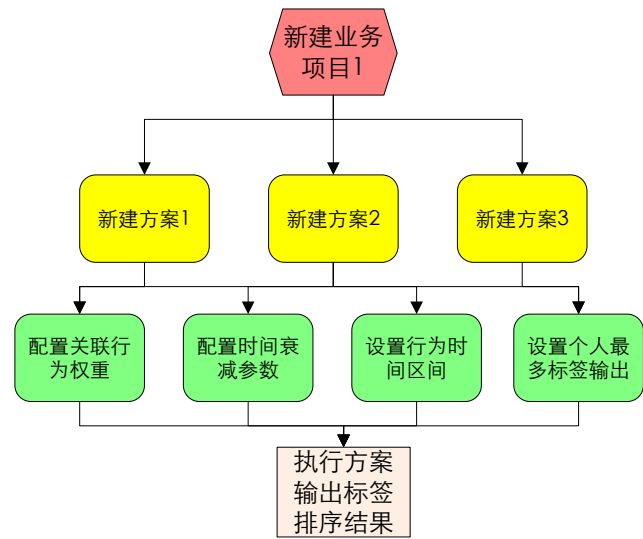


2.4 医院标签

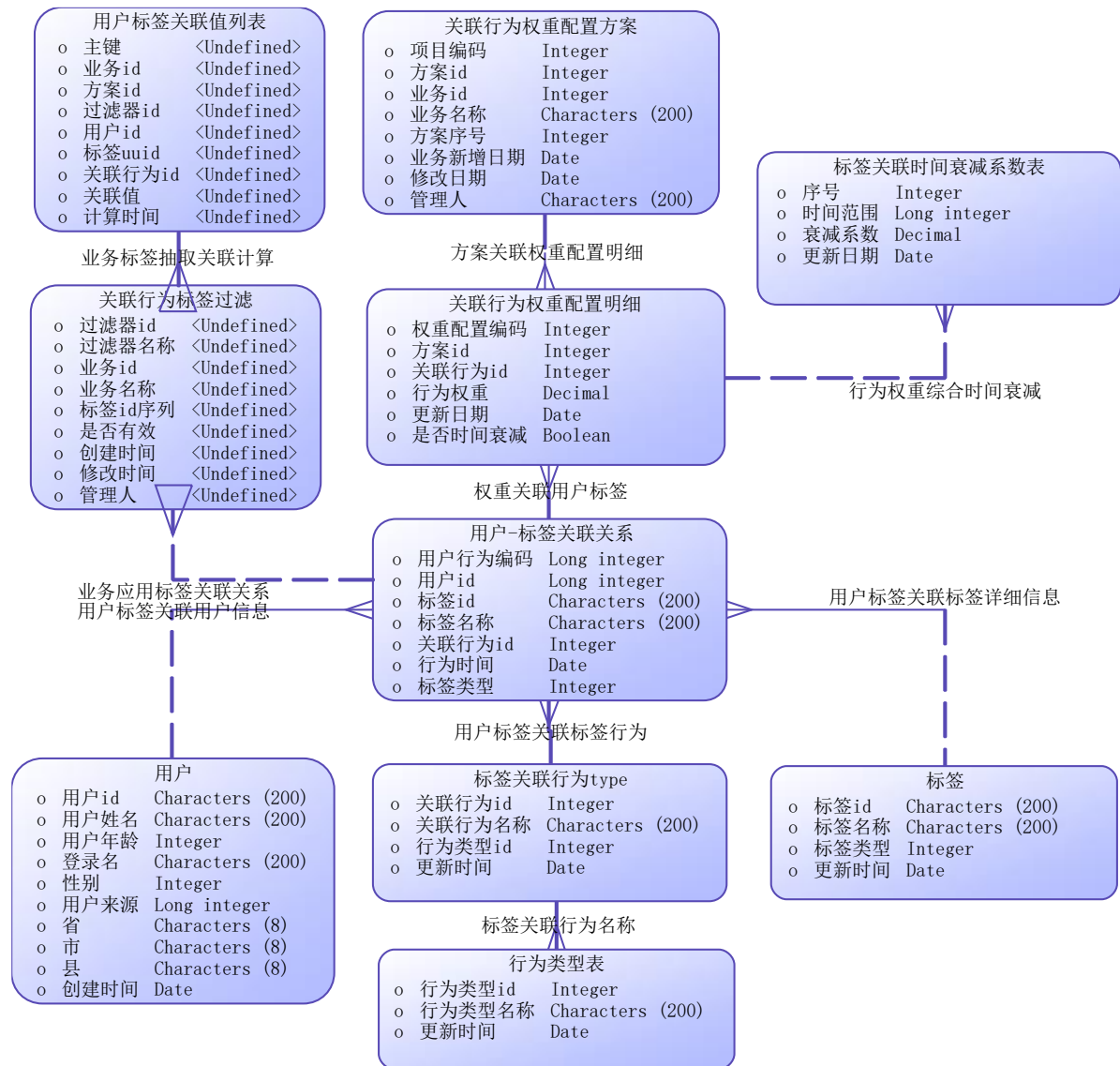


3. 应用实施流程

应用实施流程图如下：



3.1 逻辑模型



3.2 数据字典

📁 数据字典总目录：

用户就医标签数据字典					
序号	表名	表注释	表类型	责任人	建表时间
1	act_weight_plan_config	权重方案配置表	行为表	xx 数据组	2016-01
2	act_weight_plan_detail	权重配置明细表	当前表	xx 数据组	2016-01
3	user_medical_tag_relation_portrait	用户行为标签清单表	行为表	xx 数据组	2016-01
4	time_reduce_index_detail	时间衰减系数表	当前表	xx 数据组	2016-01
5	user_medical_tag_info	用户标签信息表	当前表	xx 数据组	2016-01
6	user_medical_act_info	用户行为信息表	当前表	xx 数据组	2016-01
7	user_act_info	行为类型表	当前表	xx 数据组	2016-01
8	user_info	用户信息表	当前表	xx 数据组	

3.3 权重配置

📁 医言堂优先级规则(运营 xxx 提供)：

第一优先级：医言堂内搜索标签（目前还没上线）

第二优先级：平台内搜索标签

第三优先级：医言堂内发回贴、加入医言堂圈子、医言堂内关注标签及医生

第四优先级：挂号、问诊、报到、平台关注、疾病关联

第五优先级：浏览页面、浏览医言堂帖子、浏览圈子、浏览文章

📁 基于优先级规则配置权重（建议配置）：

id	行为类型	直接/间接	权重配置
0	用户挂号自填疾病	1	0.7
1	用户挂号评价填写疾病	1	0.7
2	用户挂号医生擅长疾病	0	0.65
3	用户问诊医生诊断疾病	1	0.7
4	用户问诊评价填写疾病	1	0.7
5	用户问诊医生擅长疾病	0	0.65
6	用户报到所患疾病	1	0.7
7	用户报到医生擅长疾病	0	0.65
8	用户关注疾病	1	0.7
9	用户关注医生擅长疾病	0	0.65
10	用户关注医言堂医生擅长疾病	0	0.75
11	用户浏览专家文章对应疾病	0	0.55
12	用户浏览媒体聚焦对应疾病	0	0.55
13	用户搜索疾病标签	1	0.9
14	用户浏览圈子对应标签	0	0.6
15	用户加入圈子对应标签	1	0.8
16	用户挂号医生对应科室	1	0.7
17	用户问诊医生对应科室	1	0.7
18	用户报到医生对应科室	1	0.7
19	用户关注科室	1	0.7
20	用户关注医生对应科室	0	0.65
41	用户搜索医生标签	1	0.9
42	用户搜索医院标签	1	0.9
43	用户医言堂发回帖子对应主题标签	0	0.8
44	用户医言堂浏览帖子对应主题标签	0	0.6
45	用户医言堂关注标签	1	0.8

### 3.4 时间衰减系数

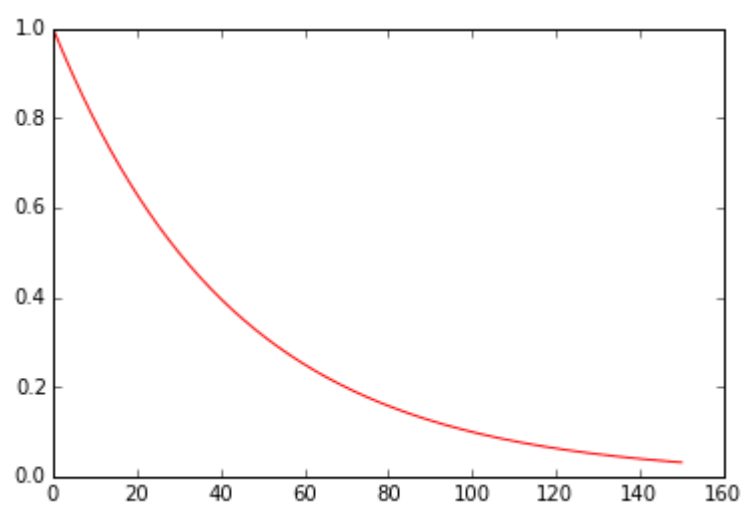
时间衰减是指用户的行为会随着时间的过去，历史行为和当前的相关性不断的减弱，其相关系数呈现指数型衰减。 $f(t)$ 是随着时间  $t$  的增长而衰减，其原来行为信息衰减为：

$$f(t) = e^{-\alpha t}$$

$\alpha$ 为衰减常数，通过回归可计算得出。如：指定 30 天前的行为信息和当前的相关度为 0.5，即  $0.5 = e^{-\alpha \cdot 30}$ ，求得  $\alpha = 0.0231$ 。

行为信息衰减函数曲线为：





### 3.5 执行结果

🚦 执行代码 (Hive) ——取一年内用户标签:

```
SELECT user_id,portrait_id,portrait_name,tag_type,
        ROW_NUMBER () OVER (PARTITION BY user_id ORDER BY weight )
weight_order
FROM (
    SELECT
pp1.user_id,pp1.portrait_id,pp1.portrait_name,pp1.tag_type,
        SUM(act_weight_index) weight
    FROM (
        SELECT
p1.user_id,p1.portrait_id,p1.portrait_name,p1.tag_type,
        EXP(datediff (TO_DATE(NOW()), p1.date_id) * (-
0.0231))
        * p2.act_weight_index
act_weight_index
        FROM gdw_dm.user_medical_tag_relation_portrait p1
    LEFT JOIN gdw_dm.act_weight_plan_detail p2 ON p1.act_tag_id =
p2.act_tag_id
        WHERE p1.date_id >= DATE_SUB (TO_DATE(NOW()), 365)
        AND p2.is_time_reduce = 1
    UNION
    SELECT
p3.user_id,p3.portrait_id,p3.portrait_name,p3.tag_type,p4.act_weight_
index
        FROM gdw_dm.user_medical_tag_relation_portrait p3
    LEFT JOIN gdw_dm.act_weight_plan_detail p4 ON p3.act_tag_id =
p4.act_tag_id
        WHERE p4.is_time_reduce = 0
```

) pp1

GROUP BY 1, 2, 3, 4

) ppp

执行结果样例:

2669 38246	子宫平滑肌瘤	1	1
2669 38343	经前期综合征	1	2
2669 38340	高催乳素血症	1	3
2669 44387	习惯性流产	1	4
2669 44423	阴道炎	1	5
2669 1f0d4589-31e0-11e6-804e-848f69fd6b70	高泌乳素血症	1	6
2669 38365	女性不育	1	7
2669 E9F2E16619FD31D7E0400A0AC864423C	子宫腺肌病	1	8
2669 26c2ed1d-31e0-11e6-804e-848f69fd6b70	卵巢早衰	1	9
2669 44478	痛经	1	10
2669 7f67f8c4-cff3-11e1-831f-5cf9dd2e7135	产科	7	11
2669 44250	宫颈炎	1	12
2669 43513	畸胎瘤	1	13
2669 37060	自然流产	1	14