Web Data策略设计文档模板说明v3.0

|  |  |
| --- | --- |
| **策略名称** | 策略的名称 |
| **所属Topic** | 策略所属的topic |
| **策略版本** | 策略的版本 |
| **负责人** |  |
| **审核人** |  |
| **文档提交日期** |  |

目录

1 背景与需求 1

1.1 背景 1

1.2 名词解释 1

1.3 需求分析 1

1.3.1 需求点说明 1

1.3.2 应用场景 1

1.3.3 效果与性能预期 2

2 外部依赖 2

2.1 数据 2

2.2 服务及工具 2

3 对外产出 3

3.1.1 数据 3

3.1.2 服务与工具 3

4 策略设计 3

5 限制与折中 3

6 思考与改进 4

6.1 已知问题 4

6.2 放弃的思路 4

6.3 可尝试的改进 4

7 策略历史及文档信息 4

7.1 历史版本及主要功能 4

7.2 策略负责人 4

7.3 文档引用 4

# 背景与需求

*本章节通过对策略的背景描述、需求的分解分析等，力求使模块接手者、策略相关PM、qa能够通过本章的阅读对本文描述的策略是什么、为什么要有这个策略、策略预期的效果是怎样的有一个概要的认识。*

## 背景

*使用文字描述策略诞生的背景和缘由。*

## 名词解释

*对文档中出现的专有名词和关键概念进行解释。是的阅读者在阅读后文时能够充分了解作者要表达的内容，避免歧义。*

## 需求分析

*这部分对策略的需求进行描述。一方面是策略开发者与PM或需求方对于需求的明确书面记录和确认，另一方面也作为QA测试设计的依据。*

### 需求点说明

*对策略所涉及的需求点进行说明。这里的需求点是策略实现方与相关需求提出方经过沟通后，分解和明确出来的。需要经过双方的确认。*

*说明的方式可以使用文字、表格罗列的方式，也可以使用UML用例图等图形化方式。*

### 应用场景

*应用场景主要是针对为下游提供应用产品的模块（如webmine的很多模块），对文档撰写时已知的策略应用进行描述。作为策略设计时的参考。而对于一些直接对产品效果负责的模块来说，可能没有下游应用方，此时可以跳过此部分*

*记录方式可以参考下面的表格：*

|  |  |  |
| --- | --- | --- |
| 应用方 | 模块 | 应用方式说明 |
|  |  |  |

### 效果与性能预期

*对策略及产出数据的效果，以及性能等给出由应用而定的预期，包括对于评估方法和标准的说明或引用。*

# 外部依赖

*给出策略所依赖的外部数据与服务。*

## 数据

*以表格形式描述策略逻辑所依赖的外部数据，与模块设计文档不同，这里更关注的是数据的用途，而模块设计文档更关注数据的格式，飞线等细节*

*描述方式可以参考下面的样表：*

|  |  |
| --- | --- |
| 数据名称 |  |
| 数据用途 |  |
| 数据提供方 |  |

## 服务及工具

*以表格形式描述策略逻辑所依赖的外部服务与工具。*

*描述方式可以参考下面的样表：*

*工具描述：*

|  |  |
| --- | --- |
| *工具名称* |  |
| 所属模块 |  |
| 工具用途 |  |

*服务描述：*

|  |  |
| --- | --- |
| 服务名称 |  |
| 功能与作用 |  |
| 提供方 |  |

# 对外产出

*描述作为策略最终产出的数据与服务。*

### 数据

*以表格形式描述策略逻辑所产出的数据、词典等。*

*描述方式可以参考下面的样表：*

|  |  |
| --- | --- |
| 数据名称 |  |
| 数据意义 |  |

### 服务与工具

*以表格形式描述封装策略逻辑的服务与工具。*

*描述方式可以参考下面的样表：*

|  |  |
| --- | --- |
| 工具/服务名称 |  |
| 功能 |  |

# 策略设计

*使用文字、图形、表格等，清晰、无歧义的描述策略的设计。*

*tips:*

1. *建议优先使用图形的方式，便于理解*
2. *需要给出策略与实现模块之间的对应关系*
3. *可以采用分层结构进行描述，先描述总体的策略框架，然后再针对各个部分的策略进行细化描述*
4. *尽量使用形式化的表述方式（如：在描述计算方式时用数学公式而不是文字），避免理解上产生歧义*

*下面给出一个策略描述的样例*

## 策略框架

## 策略细节

### url的价值

#### 目的和作用

* 价值是新收录控制模型的核心，理想情况下我们只收录有价值的链接。
* url的价值包括检索价值、调度价值、挖掘价值。其中挖掘价值目前无法做到url级别，因此目前url价值仅包含检索价值和调度价值。
* 目前的价值判断仅针对已抓url，对于未抓取的提链，暂时无法判断其价值
* 检索价值主要由bailing中的属性计算得到，调度价值主要由linkbase中的属性计算得到。

#### 形式化描述

注：

* layer,laser,inimageindex,wise是bailing中的属性
* 其余是linkbase中的属性
* level表示get和chk分级的值，目前只有线下的dlb\_select\_url\_client工具加上select的conf配置才能从linkbase中获取出该字段

|  |  |  |
| --- | --- | --- |
| 变量 | 【取值】：条件表达式 | 说明 |
| 检索价值  (index\_value) | 【1】：layer!=0 && layer!=15 && (laser<1 || 100<laser) || 0<inimageindex || 1==wise || 2==wise  【0】：其它 | * layer!=0：进行过garbage筛选 * layer!=15：非垃圾 * laser<1 || 100 <laser：非作弊 * 0<inimageindex：图片产品使用了 * 1==wise||2==wise：无线产品使用了 |
| 调度价值  (sched\_value) | 【1】：0<=level && level<=6 && crawl在最近一年之内  【0】：其它 | * 0<=level && level<=6：chk分级在0~6 * crawl在最近一年之内：一年内调度过 |

# 限制与折中

*本章内容包括：*

*1、在没有限制与折中情况下的理想策略模型是什么样的*

*2、策略设计中附加的假设和由此产生的限制条件（策略在什么条件下会失效、不适用或需要重新review等）。*

*3、策略设计中的折中考虑，和由此带来的适用性、效果方面的影响，以及review的时机等。*

# 思考与改进

## 已知问题

*记录策略当前已知而未处理的问题。作为后续升级改进的参考。*

*需要尽可能详尽的给出问题描述、影响等。*

## 放弃的思路

*记录调研中尝试过但放弃的思路。作为后续改进时的参考，一方面避免重复，另一方面从之前放弃的原因中也可以找到新的思路。*

*需要尽可能详尽的记录尝试过的内容，尝试的方法、结果，以及为何放弃该思路的原因。*

## 可尝试的改进

*列出目前考虑到但未尝试的思路，作为后续升级改进的参考。*

# 策略历史及文档信息

## 历史版本及主要功能

*给出策略的历史版本及主要改进。*

*对于一个策略框架对应于多个模块的情况，如果只涉及模块实现层面的改动，可以不用更新策略文档。但如果涉及策略层面的改动：*

1. *需要修改策略文档，保证其中所有的策略点都是最新的*
2. *在本小节中注明修改历史*
3. *在本小节中说明策略的修改与哪一个模块版本对应*

## 策略负责人

按照版本列出策略的负责rd和pm。

可以参考如下的表格描述：

|  |  |  |
| --- | --- | --- |
| 版本号 | 负责RD | 负责PM |
|  |  |  |

## 文档引用

*给出与本策略相关的文档列表。包括调研文档、调研数据、实现策略的模块的设计文档等。给出简要说明和icafe文档链接即可。*

*描述方式可以参考下面的样表：*

|  |  |  |  |
| --- | --- | --- | --- |
| 序号 | 名称 | icafe地址 | 备注 |
| 1 |  |  |  |