易淘食BI部爬虫工具使用说明

1. **开发步骤**
2. 新建统计源



1. 新建数据模型(定义表名，并在数据库中创建特定的表)



1. 在选定的表中创建表结构(定义列名，并在表中创建特定的列)



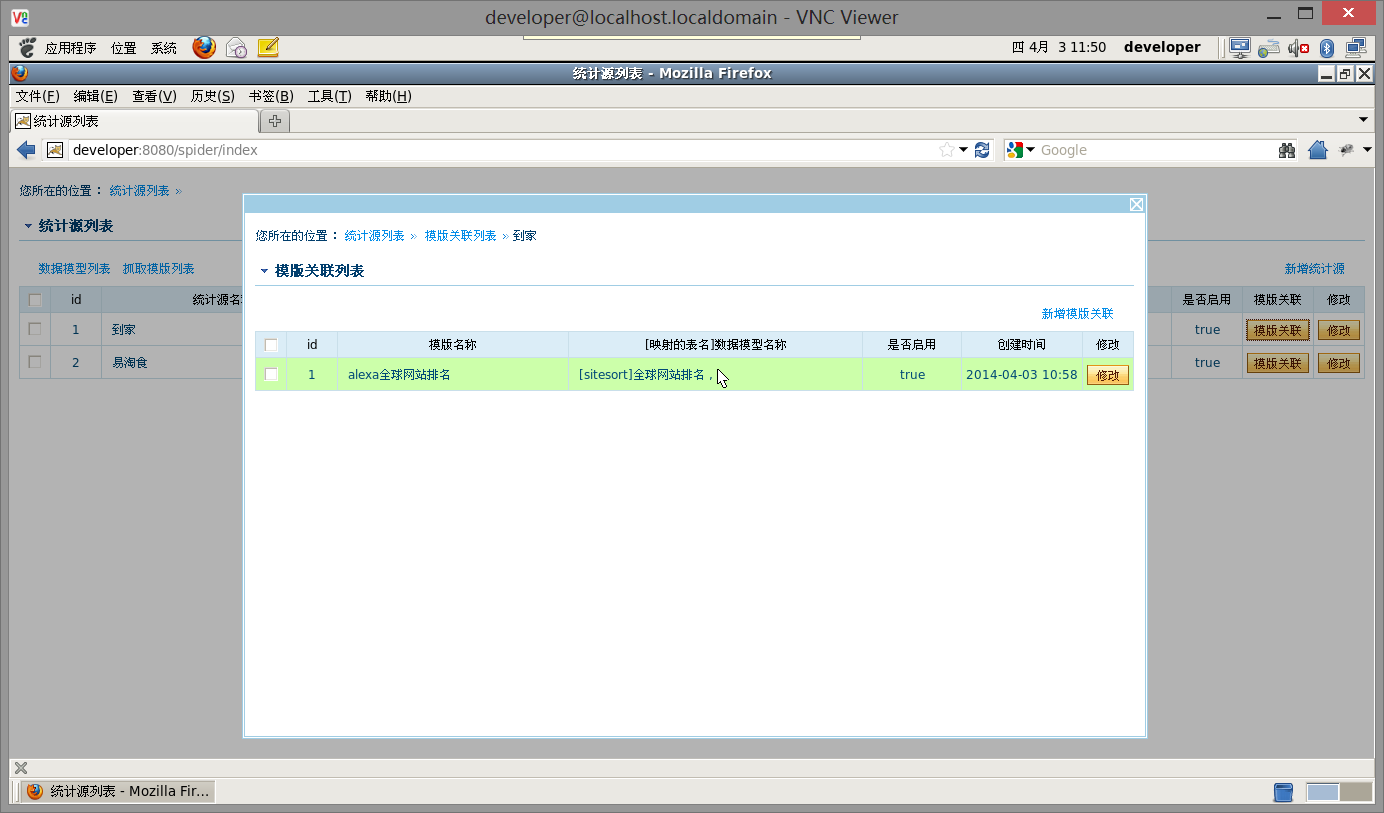
1. 定义抓取模版，并选定对应的数据模型

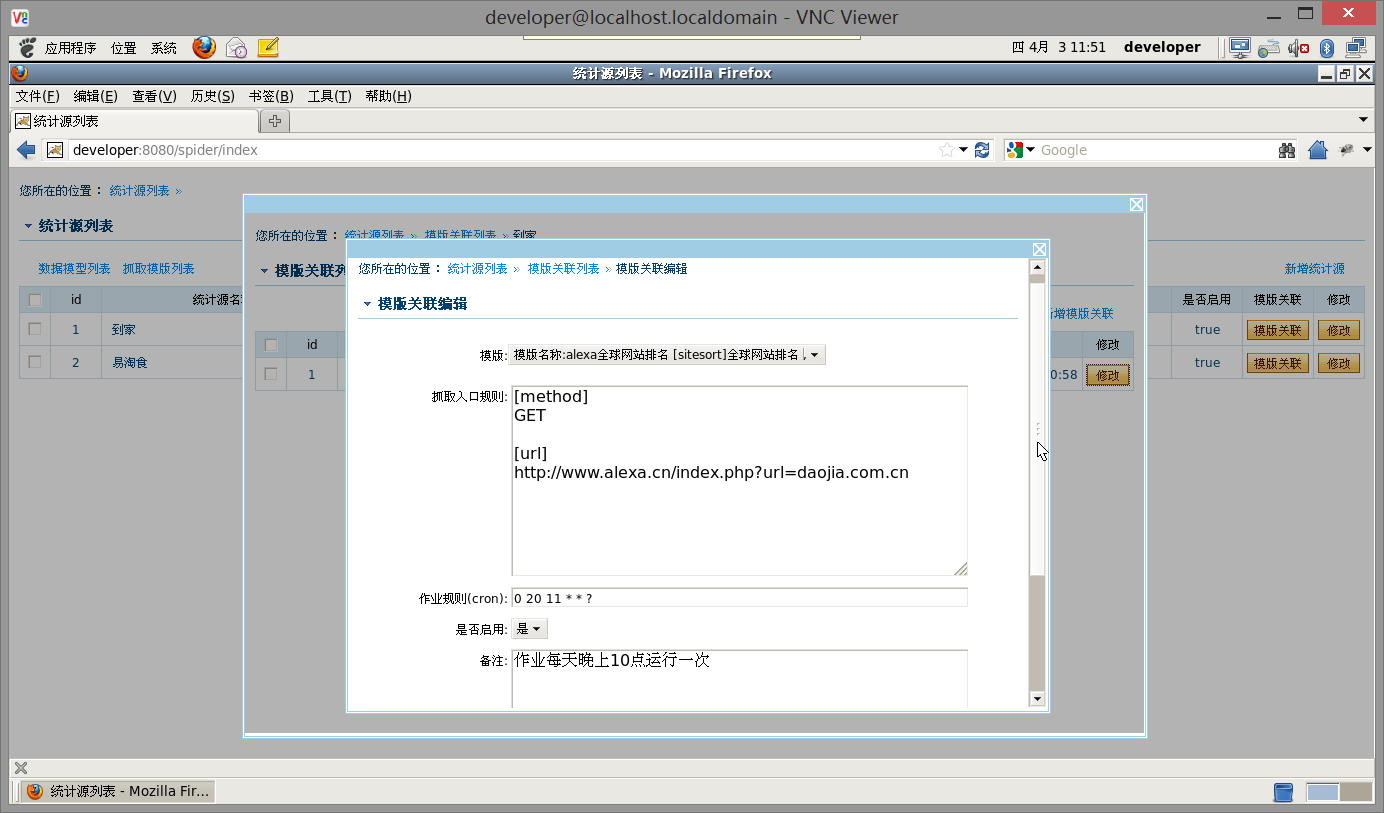


1. 配置抓取的正则模版，通过正则表达式获取需要抓取的数据



1. 统计源与抓取模版关联，并设定定时作业规则





1. **抓取的正则模版配置说明：**

规则：每次匹配只捕获一列，通过创建不同的捕获来获取不同的列

选项说明：

1. 测试url：测试正则模版需要的url源地址
2. 测试TestContent：测试正则模版需要的内容
3. 抓取类型：除了两个常用类型“[filter]过滤”、“[entry]入口”，其他都是创建数据模型的表字段
4. 正则表达式：配置的正则表达式模版
5. 按钮“Url测试”：根据“测试Url”来测试正则表达式模版
6. 按钮“Content测试”：根据“测试TestContent”来测试正则表达式模版
7. 抓取排序：抓取的顺序
8. 是否下一步抓取：如果需要递归一层，请选择“是”
9. 入口规则：如果是下一步抓取的递归入口，则需要填写入口规则（入口规则算法详见第三点），如果是其他项，则选填为正则测试中捕获url数据的算法规则（算法规则详见第四条）
10. **入口规则算法**

注意：算法的格式都必须顶头，不能留空格，内容不要换行

1. *method*算法：如果需要做一次递归抓取，则填写POST或者GET请求，格式：

[method]

GET

1. *postparams*算法：如果需要做一次递归抓取且是POST请求，则填写POST的数据项，格式：

[postparams]

cid={{array.cid}}&fid={{request.fid}}

1. *url*算法：如果需要做一次递归抓取，则填写需要抓取的url地址规则，格式：

[url]

http://www.etaoshi.com/?cid={{array.cid}}&fid={{request.fid}}

1. *content*算法：如果需要做一次内容递归（一对多关系），则需要填写content规则，格式：

[content]

{{array.cid}}

1. *response*算法：当前处理逻辑需要向下一步递归逻辑传入数据，则需要填写response规则，格式：

[response]

{{array.cid}}{{request.fid}}

1. *insertdb*算法：当前处理逻辑需要入库的请求，请填写insertdb规则，格式：

[insertdb]

Tablename({{array.cid}}{{request.cid}})

1. *header*算法：如果需要做一次递归抓取，且需要在请求url时传递http头信息，请填写header规则，格式：

[header]

((Referer={{request.url}}))((Cookie=cityid=1))

1. *ssl*算法（暂不支持）：如果是ssl请求，请填写ssl规则，格式：

[ssl]

1.cer

1. *page*算法：如果是分页请求，需要加入分页参数，请填写page规则，格式：

[page]

1-50|1

说明：其数据的含义 起始页码-结束页码|递增量

1. *uniquekey*算法：如果是入库请求时需要指定唯一键，请填写uniquekey规则，格式：

[uniquekey]

request.cityid|array.areaid

其他说明：

1. array：表示当前处理逻辑中捕获的数据数组
2. request：表示当前处理逻辑的父逻辑中传入的数据
3. **捕获Url数据的算法：算法规则参见第三点**

例子：

[method]

POST

[postparams]

url=daojia.com.cn&sig=6b35767284f726fcac68f5095df5e2ef&keyt=1396234864

1. **作业cron规则：请参见quartz的cron语法**