**抽取用XML文件撰写阅读规范**

1. **说明**

本文档是用于帮助阅读和编写类似sinaExtract.xml文件（以下简称xml文件）的说明文档。该xml文件保存了网页抽取相关的数据库信息和xpath信息。具体范例见文档第4部分。所有有上标的内容，其对应的详细说明在第5部分。

1. **Xml文件命名规范及标签树状结构**

每一个xml文件的命名规范为：网站名+Extract.xml(大小写敏感)。例如sinaExtract.xml表示的是网站sina的抽取xml文件名。

*图表 xml文件标签树形图*

1. **主要标签[1]说明**

**3.1数据库部分**

标签

路径

含义

实例

url

root/db/url

JDBC中的url一部分

<url>jdbc:mysql://localhost:3306/lin</url>

user

root/db/user

数据库帐号

<user>root</user>

password

root/db/password

数据库密码

<password>a62055974</password>

charset

root/db/charset

数据库编码

<charset>UTF-8</charset>

table

root/db/table

table下有各类子标签用以保存数据库表的各种信息

<table>……</table>

name

root/db/table/name

table下的子标签用以保存数据库表的名字

<name>sina</name>

column

root/db/table/column

table下有各类子标签用以保存数据库表的列属性

<column>……</column>

cname[2]

root/db/table/column/cname

column下的子标签用以保存数据库表的列

[2]

<cname>title</cname>

ctype

root/db/table/column/ctype

column下的子标签用以保存数据库表的列数据类型

<ctype>Varchar(100)</ctype>

isKey

root/db/table/column/isKey

column下的子标签用以标明数据库表的列是否为主键。内容只能为yes或no

<iskey>yes</iskey>

**3.2 xpath部分**

|  |  |  |  |
| --- | --- | --- | --- |
| 标签 | 路径 | 含义 | 实例 |
| rule | root/rule | 抽取规则集合 | <rule>…</rule> |
| ruleN[3] | root/ruleN | 第N个抽取规则 | <rule1>…</rule1> |
| news | root/ruleN/news | 新闻类抽取规则，其子标签保存了网页各个部分的抽取规则 | <news>…</news> |
| forum | root/ruleN/forum | 论坛类抽取规则，其子标签保存了网页各个部分的抽取规则 | <forumn>…</forumn> |

关于news，forum标签的子标签填写的一些说明：

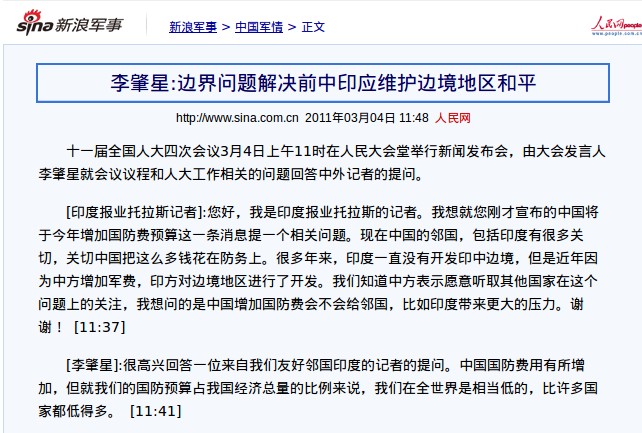
1. **news**

news标签下往往会存在一些body，publictime，title等子标签。其内容可以使用firepath插件直接从网页上获取。具体流程如下

1. 在安装有firepath的浏览器上打开一个目标网页，并启动firepath；



1. 点击图标A description...
2. 然后选择需要抽取的部分（如下图蓝框）



1. 接着在xpath会出现所需的xpath路径[4]

A description...

1. 框内的xpath就是news标签下各个子标签的内容
2. **forumn**

forum标签下必须包含mainframe标签。由于论坛往往是由一个个框组成，对于每一个框都对应一个xpath。而forum标签下的其它标签则对于每个框内内容的xpath的相对路径。

其中获得mainframe路径的方式与news类似，但需要注意的是如果获得xptah路径如图

A description...

需要对该路径进行简单加工，将其变为.//\*[@id]/tbody

对于其它标签方法也与news类似，需要注意的是对于获得的路径如图

A description...

也需要对该路径进行简单加工，删除mainframe包含的部分，变为

.//tr[1]/td[1]/div[1]/a/font

**3.3其它**

|  |  |  |  |
| --- | --- | --- | --- |
| 标签 | 路径 | 含义 | 实例 |
| urlPattern | /root/urlPattern | 被抽取网址的正则表达式 | <urlPattern><![CDATA[http://news.sina.com.cn/s/(p/)?2010-\d+-\d+/\d+.shtml]]></urlPattern> |
| appName | /root/appName | 网站名 | <appName>sina</appName> |

1. **范例**

如下为一个xml文件范例：

<?xml version=*"1.0"* encoding=*"UTF-8"*?>

<root>

<appName>sina</appName>

<db>

<url>jdbc:mysql://localhost:3306/lin</url>

<user>root</user>

<password>a62055974</password>

<charset>UTF-8</charset>

<table>

<name>sina</name>

<column>

<cname>title</cname>

<ctype>Varchar(100)</ctype>

<iskey>yes</iskey>

</column>

<column>

<cname>body</cname>

<ctype>text</ctype>

<iskey>no</iskey>

</column>

<column>

<cname>publictime</cname>

<ctype>Varchar(20)</ctype>

<iskey>no</iskey>

</column>

<column>

<cname>ref</cname>

<ctype>Varchar(50)</ctype>

<iskey>no</iskey>

</column>

</table>

</db>

<urlPattern><![CDATA[http://news.sina.com.cn/s/(p/)?2010-\d+-\d+/\d+.shtml]]></urlPattern>

<rule>

<rule1>

<news>

<title>//\*[@id='artibodyTitle']</title>

<ref>//\*[@id='art\_source']/a</ref>

<body>//\*[@id='artibody']/p</body>

<!--<body>.//\*[@id='page']/div[4]/div[1]/div[1]/div[1]</body>-->

<publictime>//\*[@id='pub\_date']</publictime>

</news>

<forum name=*"jisu"*>

<mainframe>.//\*[@id]/tbody</mainframe>

<username>.//tr[1]/td[1]/div[1]/a/font</username>

<content>.//tr/td/div/div/div/table/tbody/tr/td</content>

</forum>

</rule1>

</rule>

</root>

1. **注意事项**
   1. 所有标签大小写敏感；
   2. root/db/table/column/cname内的值必须一一对应于root/ruleN/news的各个子标签名或者root/ruleN/forum内除mainframe外的各个子标签名；
   3. N表示正整数；
   4. 路径中可能存在一些常量，需要对其进行简单加工，例如删除该常量