Cấu hình Database (sử dụng module dyna persistence): conf/database/database.properties

datasource.class=org.apache.commons.dbcp.BasicDataSource

datasource.driverClassName=com.mysql.jdbc.Driver

datasource.url=jdbc:mysql://localhost/test?characterEncoding=UTF-8

datasource.username=root

datasource.password=

Cấu hình site:

VD: site dân trí:

Cấu trúc folder:

Sites:

------dantri:

--------------configuration.xml (file config)

--------------lasttime.dat (file time crawl phục vụ việc resume)

--------------newconf.xml (file config cách crawl các trường)

1.Configuration.xml:

<?xml version=*"1.0"*?>

<?xml-stylesheet type=*"text/xsl"* href=*"configuration.xsl"*?>

<!-- Put site-specific property overrides in this file. -->

<configuration>

<property>

<name>entity.class</name>

<value>dyna#article</value>

</property>

<property>

<name>entity.id.field</name>

<value>articleID</value>

</property>

</configuration>

2 properties bắt buộc:

**entity.class** : định nghĩa kiểu thực thể, nó có thể là tên của đầy đủ của 1 class vd như: com.ant.core.entity.Article ( nếu sử dụng tên class sẽ dùng module hibernate để store data) hoặc có thể là tên của 1 bảng trong csdl (sẽ có tiền tố bắt đầu bằng dyna#) như ở đây là article

**entity.id.field:**tên trường primary key của thực thể

**Option properties:**

**entity.thumbnail.field:** tên trường lưu trữ ảnh thumbnail (phục vụ cho các thực thể muốn lấy thumbnail)

Ngoài ra còn thêm các thuộc tính khác override các thuộc tính trong file conf/crawler.xml (tuy nhiên h chưa cày đặt)

Vd: ở trên sử dụng bảng article

**article**

| **Column** | **Type** | **Null** | **Default** |
| --- | --- | --- | --- |
| articleID | int(11) | No |  |
| categoryID | int(11) | No |  |
| title | varchar(500) | Yes | *NULL* |
| content | text | No |  |
| createDate | timestamp | Yes | *NULL* |
| status | tinyint(4) | Yes | *NULL* |
| hot | tinyint(4) | Yes | *NULL* |

2. newconf.xml

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

<entityConf xsi:noNamespaceSchemaLocation=*"../../conf/entityconf.xsd"*

xmlns:xsi=*"http://www.w3.org/2001/XMLSchema-instance"*>

<entityFields>

<rssSite>

<mappingField rssField=*"title"* entityField=*"title"*>

<filter>…</filter>

<filter>…</filter>

</mappingField>

</rssSite>

<listSite>

<itemXpath></itemXpath>

<field xpath=*""* name=*""*>

<filter>…</filter>

</field>

</listSite>

<detailSite>

<field name=*"content"*

xpath=*"//div[@id='ctl00\_IDContent\_Tin\_Chi\_Tiet']/div[@id='ctl00\_IDContent\_ctl00\_divContent']/div[@class='fon34 mt3 mr2 fon43']/\*[not(self::input)]"*>

<filter>…</filter>

<filter>…</filter>

</field>

</detailSite>

</entityFields>

<categories mappingField=*"categoryID"*>

<category id=*"39"*>

http://dantri.com.vn/otoxemay.rss

</category>

<category id=*"44"*>

http://dantri.com.vn/skphapluat.rss

</category>

<category id=*"43"*>

http://dantri.com.vn/giaoduc-khuyenhoc.rss

http://dantri.com.vn/john-linh.rss

http://dantri.com.vn/guong-sang.rss

http://dantri.com.vn/duhoc.rss

</category>

</categories>

</entityConf>

Có 2 loại website RSS và non-RSS

Site RSS sẽ có backend crawl riêng và sử dụng thẻ <rssSite> để cấu hình một số trường

RSS mapping các trường trong đối tượng item với các trường trong ${ **entity.class** }

Site non-RSS sử dụng backend crawl: com.ant.crawler.core.ListSiteCrawler cấu hình như sau:

<entityConf xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"

xsi:noNamespaceSchemaLocation="../../conf/newsconf.xsd" backend="com.ant.crawler.core.ListSiteCrawler">

Và site non-RSS sẽ dùng thẻ <listSite> để định nghĩa cách lấy các item từ site list.

Trong thẻ categories có thuộc tính mappingField, đây là thuộc tính tên của trường category trong thực thể.

Run chương trình:

Chạy class com.ant.crawler.core.MainExecutor

Với tham số memory cho máy máy ảo tối thiểu là -Xms512m -Xmx1024m