**大作业报告之二---填充构架中的技术点**

By --- 吴辉

我们组的整个项目，采用开源的爬虫框架WebMagic。构架中的技术点及其分配如下：

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 技术分工  架构层 | 使用技术 | 负责人 | 处理过程 | 预期效果 |
| 数据获取 | WebMagic开源框架  该框架分为Spider(整个爬虫的调度框架)、Downloader(页面下载)、PageProcessor(链接提取和页面分析)、Scheduler(URL管理)、Pipeline(离线分析和持久化)几部分，我们获取了该框架的API之后，即可调用Spider类进行调度抓取网页信息 | 吴 辉  谢年韬 | 具体demo见下述 | 通过webMagic获取到CSDN、博客园等相关技术博客网站的编程语言（C、C++、java、python等）、大数据技术（Hadoop、openstack等）等技术博客的博文数据，保存在数据库中。 |
| 数据组织与存储 | SQLite存储  Json组织形式  Xml组织形式 | 谢年韬  宣梦雨 | 将第一步获取到的数据通过SQLite存储到数据库中，通过分析数据库中的数据类型和内容，组织数据形式，存入持久化对象中。比如，从CDSN博客中获取到的关于python分类的博文，我们在数据库中加入两个字段—FromUrl(代表博文来源，一般是网址形式)、BlogTag(标签，即python) | 数据通过SQLite存储在关系型数据库中，并可以通过持久层对象访问。 |
| 数据展示 | Html技术  Css技术  Jsp技术 | 吴 辉  宣梦雨 | 通过JavaWeb框架从数据库中获取数据并通过web页面展示 | 和一般的博客分类网站效果类似 |

通过webMagic获取数据demo:

代码：

package com.spider;

import java.util.List;

import us.codecraft.webmagic.Page;

import us.codecraft.webmagic.Site;

import us.codecraft.webmagic.Spider;

import us.codecraft.webmagic.pipeline.ConsolePipeline;

import us.codecraft.webmagic.processor.PageProcessor;

public class OschinaBlogPageProcesser implements PageProcessor {

private Site site = Site.me().setDomain("my.oschina.net")

.addStartUrl("http://my.oschina.net/flashsword/blog/123505?fromerr=jN5AybQL");

@Override

public void process(Page page) {

List<String> links = page.getHtml().links().regex("http://my\\.oschina\\.net/flashsword/blog/123505?fromerr=jN5AybQL\\d+").all();

page.addTargetRequests(links);

System.out.println(page.getHtml().xpath("//title"));

page.putField("title", page.getHtml().xpath("//title").toString());

page.putField("content", page.getHtml().xpath("//div[@class='BlogContent']/a/text()").all());

page.putField("tags",page.getHtml().xpath("//div[@class='BlogTags']/a/text()").all());

}

@Override

public Site getSite() {

return site;

}

public static void main(String[] args) {

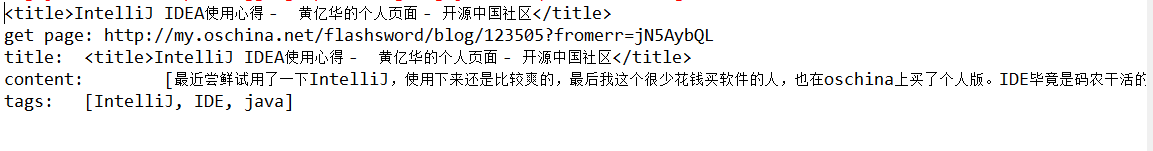
Spider.create(new OschinaBlogPageProcesser())

.pipeline(new ConsolePipeline()).run();

}

}

实现效果：在控制台打印如下信息：



原网页信息：

