.NET设计与架构实验报告

53140808 刘世奇

目录

**摘要1**

**一．简介**1

1网络爬虫的构成及分类1

2网络爬虫的工作原理2

3网络爬虫的搜索策略3

**二．研究方法3**

**三．数据分析和结果7**

**四．结论9**

**附录1：9**

**附录2：11**

摘要：

使用了Selenium驱动进行了爬取并用c#进行数据处理。

一、简介：

1.网络爬虫的构成及分类

网络爬虫又被称为做网络蜘蛛、网络机器人，主要用于网络资源的收集工作。在进行网络舆情分析时，首要获取舆情信息内容，这就需要用到网络爬虫(蜘蛛程序)这个工具，它是一个能自动提取网页内容的程序，通过搜索引擎从互联网上爬取网页地址并抓取相应的网页内容，是搜索引擎(Search Engine)的重要组成部分。

一个典型的网络爬虫主要组成部分如下：

1) URL 链接库，主要用于存放爬取网页链接。

2)文档内容模块，主要用于存取从 Web 中下载的网页内容。

3)文档解析模块，用于解析下载文档中的网页内容，如解析 PDF，Word，HTML 等。

4)存储文档的元数据以及内容的库。

5)规范化 URL 模块，用于把 URL 转成标准的格式。

6)URL 过滤器，主要用于过滤掉不需要的 URL。

上述模块的设计与实现，主要是确定爬取的内容以及爬去的范围。最简单的例子是从一个已知的站点抓取一些网页，这个爬虫用少量代码就可以完成。然而在实际互联网应用中，可能会碰到爬去大量内容需求，就需要设计一个较为复杂的爬虫，这个爬虫就是N个应用的组成，并且难点是基于分布式的。

2.网络爬虫的工作原理

传统网路爬虫的工作原理是，首先选择初始 URL，并获得初始网页的域名或 IP 地址，然后在抓取网页时，不断从当前页面上获取新的 URL 放入候选队列，直到满足停止条件。

聚焦爬虫(主题驱动爬虫)不同于传统爬虫，其工作流程比较复杂，首先需要过滤掉跟主题不相关的链接，只保留有用的链接并将其放入候选 URL 队列。然后，根据搜索策略从候选队列中选择下一个要抓取的网页链接，并重复上述过程，直到满足终止条件为止。与此同时，将所有爬取的网页内容保存起来，并进行过滤、分析、建立索引等以便进行性检索和查询。

总体来讲，网络爬虫主要有如下两个阶段：

第一阶段，URL 库初始化然后开始爬取。

第二阶段，爬虫读取没有访问过的 URL，来确定它的工作范围。 其中，对于所要抓取的 URL 链接，进行以下步骤：

1)获取 URL 链接。

2)解析内容，获取 URL 及相关数据。

3)存储有价值的数据。

4)对新抓取的 URL 进行规范化

5)过滤掉不相关的 URL。

6)将要抓取的 URL 更新到 URL 库中。

7)重复步骤2，直到终止条件为止。

3.网络爬虫的搜索策略

目前，比较常见的网络爬虫搜索策略有以下三种：

1)广度优先搜索策略。其主要思想是，由根节点开始，首先遍历当前层次的搜索，然后才进行下一层的搜索，依次类推逐层的搜索。这种策略多用在主题爬虫上，因为越是与初始 URL 距离近的网页，其具有的主题相关性越大。

2)深度优先搜索策略。这种策略的主要思想是，从根节点出发找出叶子节点，以此类推。在一个网页中，选择一个超链接，被链接的网页将执行深度优先搜索，形成单独的一条搜索链，当没有其他超链接时，搜索结束。

3)最佳优先搜索策略。该策略通过计算 URL 描述文本与目标网页的相似度，或者与主题的相关性，根据所设定的阈值选出有效 URL 进行抓取。

二、研究方法：

使用了Selenium框架驱动Chrome浏览器进行爬取，安全不易封禁。

三、数据分析和结果：

分离了微博的时间和内容，以json格式储存

注意：运行需要Selenium，以及在Crawler.cs中设置自己的微博用户名和密码。

Git: henofdante/Selenium

四、结论：

通过对于结果的分析可以得出，该程序可以较好的完成任务。

不足：

在控制台执行程序，没有较好的用户界面，此项功能可以进一步完善。

附录1：

程序代码：

//Crawler.cs

using System;

using System.Collections.Generic;

using System.Linq;

using System.Text;

using System.Threading.Tasks;

using System.Threading;

using OpenQA.Selenium;

using OpenQA.Selenium.Chrome;

namespace SeleniumTest

{

class Crawler

{

public delegate bool Cycle(IWebElement root);

private IWebDriver \_driver;

private Cycle \_cycle;

private const string User = "daliansunziyu";

private const string Username = "17649818647";

private const string Password = "qwerty123456.";

private const string Url = "https://weibo.cn/";

private const string LoginUrl = "https://passport.weibo.cn/signin/Login";

private void Init()

{

\_driver = new ChromeDriver();

}

private void Login()

{

\_driver.Navigate().GoToUrl(LoginUrl);

var userName = \_driver.FindElement(By.Id("loginName"));

userName.SendKeys(Username);

var password = \_driver.FindElement(By.Id("loginPassword"));

password.SendKeys(Password);

\_driver.FindElement(By.Id("loginAction")).Click();

Thread.Sleep(7500);

}

public void Start()

{

Init();

Login();

int page = 1;

String rawUrl = Url + User + "?page=";

int maxPage = 1;

Console.WriteLine("Login complete, navigating to " + Url + User);

\_driver.Navigate().GoToUrl(Url + User);

var rawMaxPage = \_driver.FindElement(By.XPath("//\*[@id=\"pagelist\"]/form/div/input[1]"));

maxPage = int.Parse(rawMaxPage.GetAttribute("value"));

while (page != maxPage)

{

\_driver.Navigate().GoToUrl(rawUrl + page++);

IWebElement root = \_driver.FindElement(By.TagName("html"));

if (!\_cycle(root))

{

page--;

}

}

}

public void Register(Cycle cycle)

{

\_cycle += cycle;

}

public static void Main()

{

Parser parser = new Parser();

Crawler crawler = new Crawler();

crawler.Register(parser.Parse);

crawler.Start();

}

}

}

//Parser.cs

using System;

using System.Collections.Generic;

using System.Linq;

using System.Text;

using System.Threading;

using System.Threading.Tasks;

using OpenQA.Selenium;

using OpenQA.Selenium.Chrome;

namespace SeleniumTest

{

class Parser

{

private Pipeline \_pipeline;

public Parser()

{

\_pipeline = new Pipeline();

}

public bool Parse(IWebElement element)

{

try

{

var elements = element.FindElements(By.ClassName("c"));

for (var i = 1; i < elements.Count - 2; ++i)

{

var time = elements[i].FindElement(By.XPath(".//span[@class='ct']")).Text;

var content = elements[i].FindElement(By.XPath(".")).Text;

content = content.Replace(time, "");

\_pipeline.Push(time, content);

}

element.FindElement(By.LinkText("下页"));

}

catch (NoSuchElementException)

{

return false;

}

return true;

}

public static void Out(String s)

{

Console.WriteLine(s);

}

}

}

//pipeline.cs

using System;

using System.Collections.Generic;

using System.IO;

using System.Linq;

using System.Text;

using System.Threading.Tasks;

namespace SeleniumTest

{

class Pipeline

{

private const string path = "d:\\final.txt";

public void Push(string time, string text)

{

string txt = "{'time':'" + time + "', 'content':'" + text + "'}";

FileStream fs = new FileStream(path, FileMode.Append);

StreamWriter sw = new StreamWriter(fs);

sw.Write(txt);

sw.Flush();

sw.Close();

fs.Close();

Console.WriteLine("file saved:" + txt);

}

}

}

附录2：数据节选

{

'time':'06月18日 09:26 来自360安全浏览器',

'content':'[置顶][怒]#独家视频#【大连凯丹广场上演暴力打人砸店事件 如此营商环境？】6月16日13时许，在大连凯丹广场发生打人砸店事件！女店主被邻店铺夫妻俩殴打至昏迷受伤，被打女店主曾多次遭对方威胁“不准卖重样货品”，被打昏迷后遭二次打砸！我市正打造一流营商环境，如此欺行霸市行为应给予严惩！你怎么看？ 全文 赞[21] 转发[46] 评论[101] 收藏 '

}{

'time':'34分钟前 来自360安全浏览器',

'content':'[吃惊]刚刚，据网友爆料，大连恒大帝景发生火情，具体情况正在进一步了解中。 秒拍视频 赞[1] 转发[1] 评论[5] 收藏 '

}{

'time':'今天 12:45 来自360安全浏览器',

'content':'[允悲]【男主播扮小天鹅，骚扰路人被城管抬走】6月19日下午，江苏南京一名身着芭蕾舞服装的男子在夫子庙搞网络直播，并对路人摆出不雅动作。遭多人投诉后，男子不听劝阻与城管发生肢体冲突，最后被强行带上警车。(视频来源：当地相关人士提供)澎湃新闻的秒拍视频 赞[3] 转发[2] 评论[20] 收藏 '

}{

'time':'今天 12:38 来自360安全浏览器',

'content':'[允悲]其实这才是《深夜食堂》的正确打开方式……竟然毫无违和感！[笑而不语] 秒拍视频 赞[12] 转发[13] 评论[6] 收藏 '

}{

'time':'今天 11:54 来自360安全浏览器',

'content':'[怒骂]遇到这样打车的，你有招吗？不光打车不给钱，还赖在车上不走了，出租车司机真是气不打一处来啊~~ 秒拍视频 赞[10] 转发[11] 评论[36] 收藏 '

}{

'time':'今天 11:09 来自360安全浏览器',

'content':'转发了 大连岭南体育中心 的微博:发布了头条文章：《送票啦|2017大连奥林匹克日城市炫彩乐跑》[哆啦A梦开心]又一场盛大的路跑活动，赶快加入我们吧！ 送票啦|2017大连奥林匹克日城市炫彩乐跑 赞[0] 原文转发[5] 原文评论[2]

转发理由:[色][色][色]受啊~~泳池狂欢，劲爆DJ，搞笑快闪，幸运抽奖！这个大夏天肆意奔跑，赶紧抢票，请叫我雷锋！ 赞[1] 转发[1] 评论[2] 收藏 '

}{

'time':'今天 10:53 来自360安全浏览器',

'content':'转发了 大连市二手车 的微博:#大连即时小新闻# 1.机动车上牌选号由10选1扩大为20选1

2.二手车过户新车主可以使用保留的原号，原车号使用一年以上，6月20日前的自动保留半年，6月20日以后自动保留一年，二手车可以用原来的车牌号@大连54青年 @辽沈晚报 @头条新闻 @大连微博广场 @大连24小时 @大连成 @大连大小事儿 @大连日报 全文 [组图共3张]

原图 赞[2] 原文转发[9] 原文评论[8]

转发理由:[赞]好啊~~~ 赞[4] 转发[2] 评论[11] 收藏 '

}{

'time':'今天 10:52 来自360安全浏览器',

'content':'转发了 JongSuk-Queen 的微博:真人真事，大家帮忙转发一下，妹妹已经走丢36个小时了！家里人急死了，大家帮忙转发一下，朋友家妹妹大家走路时帮我留意一下@大连54青年 @大连大小事儿 13614260108，谢谢！[祈祷][祈祷] 大连·金州区 [组图共4张]

原图 赞[0] 原文转发[9] 原文评论[1]

转发理由:[话筒]帮找！ 赞[2] 转发[7] 评论[3] 收藏 '

}{

'time':'今天 10:31 来自360安全浏览器',

'content':'转发了 大连网警巡查执法 的微博:#网络安全法# [话筒]【贯彻《网络安全法》，让我们共同携手“共筑网络安全，共享网络文明”】《中华人民共和国网络安全法》是为保障网络安全，维护网络空间主权和国家安全、社会公共利益，保护公民、法人和其他组织的合法权益，促进经济社会信息化健康发展而制定的。 [组图共9张]

原图 赞[2] 原文转发[32] 原文评论[9]

转发理由:[话筒]网络安全，不容小觑！共同携手“共筑网络安全，共享网络文明 ！ 赞[1] 转发[1] 评论[0] 收藏 '

}{

'time':'今天 10:27 来自360安全浏览器',

'content':'[吃惊]【奇葩！大连52岁设计师酒后“无影脚”踢碎银行大门被拘】大连一52岁男子宋某是一名设计师，6月18日21时，在西安路一家银行ATM前，因喝酒误操作致银行卡被吞！恼羞成怒的宋某暴跳如雷，爆踹银行大门致玻璃破碎，并打砸ATM机，触动警报，被赶到的辖区民警抓获，目前该男子被警方拘留10日。（据大 全文 赞[3] 转发[1] 评论[12] 收藏 '

}{

'time':'今天 08:38 来自360安全浏览器',

'content':'[怒]【大连男子当众侮辱环卫工：你是垃圾 信不信我撞死你】18日晚间，在大连甘井子区千山路与苍山路交会路口，62岁的老环卫工黄占昌在捡拾马路上的垃圾时被一辆SUV撞倒。面对倒地的黄占昌，车上一名年长的男子对着环卫工人发飙。“ 你是垃圾”、“环卫工怎么了？信不信我撞死你……”不少现场市民气愤 全文 [组图共3张]

原图 赞[15] 转发[7] 评论[65] 收藏 '

}{

'time':'今天 08:34 来自微博搜索',

'content':'转发了 怕妻诺夫司机 的微博:#大连突发# 东港两台红色别克，怼了。这次选择了不知名的星光域长江东路路口。我曾呼吁东港商务区这几个路口安装红绿灯，呼吁有二年了吧，呵呵呵，红绿灯还在来的路上。等死人了有关部门就重视了。 [组图共7张]

原图 赞[3] 原文转发[11] 原文评论[11]

转发理由:[衰]每天一撞啊~~ 赞[3] 转发[3] 评论[21] 收藏 '

}{

'time':'今天 08:33 来自360安全浏览器',

'content':'[话筒]帮找~~大伙留一下

原图 赞[1] 转发[3] 评论[1] 收藏 '

}{

'time':'今天 08:26 来自荣耀7 世界有点不同',

'content':'转发了 瓦房店新鲜事 的微博:#大连身边事# 寻找亲生父母。2017年1月9日20时许，在瓦房店谢屯镇莲花山村郑屯路边，葛女士回家途中在路边发现一个月大小的男婴，于是将其带回家抚养，至今未找到该男婴的亲生父母及家属，于2017年6月14日到谢屯派出所报警求助公安机关帮助查找该男婴的亲生父母。如有知情者请与谢屯派出所民警联系。

全文 [组图共6张]

原图 赞[4] 原文转发[18] 原文评论[18]

转发理由:[悲伤] 赞[0] 转发[4] 评论[9] 收藏 '

}{

'time':'今天 06:25 来自荣耀7 世界有点不同',

'content':'转发了 逍遥逗点 的微博:今天是某小学第二天办理入学日！昨天早上5:10我们到指定地点排队，结果6:30发号才知道一天就发60个号，我们很荣幸的排在60之外！就有了今天也需要排队！早上下着小雨，我父亲3:40到了，先拿到一个模拟号，50几。到6:30换真正的号牌！3:40才拿到50以外……我真心想知道中国教育，大连教育为何要如此为难 全文 赞[6] 原文转发[12] 原文评论[63]

转发理由:[悲伤]雨中！可怜天下父母心！ 赞[10] 转发[7] 评论[22] 收藏 '

}{

'time':'今天 06:18 来自荣耀7 世界有点不同',

'content':'[偷笑]亲，打个针而已，至于吗？ 秒拍视频 赞[15] 转发[27] 评论[32] 收藏 '

}{

'time':'今天 05:59 来自荣耀7 世界有点不同',

'content':'[酷]大连终于下雨了～ [组图共4张]

原图 赞[22] 转发[4] 评论[16] 收藏 '

}{

'time':'06月19日 21:26 来自荣耀7 世界有点不同',

'content':'转发了 ricky\_小宝 的微博:4月10号在甘井子区柏林紫郡附近的派出所办理更换身份证业务，后期该派出所搬迁并未通知新办公所在地，所以5月份经多方打听找到椒金山派出所办公地点，因为6月19日要登记注册结婚所以5月到6月多次来到椒金山派出所询问身份证进度，每次得到的答复就是等几天，今日6月19日一早在次来到椒金山派出所得到的 全文 赞[9] 原文转发[19] 原文评论[2]

转发理由:[挖鼻]@大连户口身份证 看看这个～ 赞[4] 转发[5] 评论[15] 收藏 '

}{

'time':'06月19日 21:13 来自360安全浏览器',

'content':'[衰]【大学生网上“改成绩”被骗两千元，找“网警”维权再被骗八千元】近日，南京一大学生周某在网上发现一个"大学生修改成绩"的网页，他联系了所谓的黑客，并分两次转账两千元，但对方收钱后就将他拉黑，成绩却没有任何改变。周某发现自己上当后，没通过正常途径报警，而是在网上搜索了所谓的"网络警 全文 赞[3] 转发[5] 评论[10] 收藏 '

}{

'time':'06月19日 21:03 来自360安全浏览器',

'content':'[挖鼻]【革镇堡刘书记来播报达沃斯！网友：影响20万人的矿坑啥时候治理?】今天，一段《革镇堡刘书记来播报达沃斯！这大连话说的血待亲！》在网络上疯传，我们暂且不问革镇堡到底有没有这位刘书记，就想问问，影响20多万居民，位于革镇堡周边的矿坑啥时候治理？锃明瓦亮的幸福感不是只靠嘴说的！ [组图共2张]

原图 赞[8] 转发[3] 评论[6] 收藏 '

}{

'time':'06月19日 20:51 来自360安全浏览器',

'content':'[挖鼻]【革镇堡刘书记来播报达沃斯！网友：影响20万人的矿坑啥时候治理?】今天，一段《革镇堡刘书记来播报达沃斯！这大连话说的血待亲！》在网络上疯传，我们暂且不问革镇堡到底有没有这位刘书记，就想问问，影响20多万居民，位于革镇堡周边的矿坑啥时候治理？锃明瓦亮的幸福感不是只靠嘴说的！ 全文 赞[8] 转发[9] 评论[16] 收藏 '

}{

'time':'06月19日 20:42 来自360安全浏览器',

'content':'转发了 大连54青年 的微博:[吃惊]【大连一小伙遭虐打凌辱扔进海 贷款4千“赎身”】6月10日，在大连星海湾浴场，大连小伙孙某遭6名年轻男子（最大20岁）凌辱殴打，被推扔进大海强灌海水20多分钟。随后被逼贷款4000元“赎身”。目前，大连星海湾公安分局已经将涉嫌抢劫的5名嫌犯（1人在逃）刑事拘留！PS：交友须谨慎！（据大连电视 全文 赞[78] 原文转发[34] 原文评论[235]

转发理由:[心]//@这是仙乐吖: 被害人是我现实生活中的一个好哥哥 对我帮助很多认识他的人都知道为人很讲究。 这几个人我也知道那展毓卢世祥都是社会毒瘤我哥对他们很好还这样 这个结果真的自作自受。这次希望你们也能好好反省 别在祸害社会了。我也希望我哥不会为了这个事有什么大的伤害 你是最坚强的 赞[10] 转发[0] 评论[7] 收藏 '

}