link: null
title: 珠峰架构师成长计划
description: 有新的数据的时候
keywords: null
author: null
date: null
publisher: 珠峰架构师成长计划
stats: paragraph=19 sentences=71, words=390

1. 爬取内容

1. http/axios 等爬取API接口

```
let axios = require('axios');
axios.get('https://follow-api-ms.juejin.im/v1/getUserFollowInfo?uid=551d6923e4b0cd5b623f54da&src=web')
    .then(res => console.log(res.data))
```

2. superagent/request/crawl爬取HTML页面

```
let request = require('request');
request('https://juejin.im/tag/%E5%89%8D%E7%AB%AF', (err, response, body) => {
    let regexp = /class="title" data-v-\w+>(.+?)/g;
    let titles = [];
    body.replace(regexp, (matched, title) => {
        titles.push(title);
    });
    console.log(titles);
});
```

3. 使用puppeteer控制chromium

- puppeteer是Chrome团队开发的一个node库
- 可以通过api来控制浏览器的行为，比如点击，跳转，刷新，在控制台执行js脚本等等
- 通过这个工具可以用来写爬虫，自动签到，网页截图，生成pdf，自动化测试等

```
(async () => {
     const browser = await puppeteer.launch();
     const page = await browser.newPage();
     await page.goto('https://www.baidu.com');
     await page.screenshot({ path: 'baidu.png' });
     await browser.close();
})();
```

```
const puppeteer=require('puppeteer');
const fs=require('fs');
(async function () {
        const browser=await puppeteer.launch({headless:false});
        const page=await browser.newPage();
        await page.goto('https://juejin.im/tag/%E5%89%8D%E7%AB%AF', {
waitUntil: 'networkidle2'
        });
        await page.waitFor(500);
        let comments = await page.$eval('a.title', els => {
return els.map(item => item.innerText);
        });
        fs.writeFileSync('comments.txt',comments.join('\r\n'),'utf8');
        await browser.close();
})();
```

```
const puppeteer=require('puppeteer');
(async function () {
    const browser=await puppeteer.launch({headless:false});
    let page = await browser.newPage();
    await page.setJavaScriptEnabled(true);
    await page.goto("https://www.jd.com/");
    const searchInput = await page.$("#key");
    await searchInput.focus();
    await page.keyboard.type("手机");
    const searchBtn = await page.$(".button");
    await searchBtn.click();
    await page.waitForSelector('.gl-item');
    const links = await page.$eval('.gl-item > .gl-i-wrap > .p-img > a', links => {
        return links.map(a => {
            return {
                href: a.href.trim(),
                title: a.title
            }
        });
    });
    page.close();
    const aTags = links.splice(0, 1);
    for (var i = 0; i < aTags.length; i++) {
        page=await browser.newPage();
        page.setJavaScriptEnabled(true);
        await page.setViewport({
            width: 1920,
            height: 1080
        });
        var a = aTags[i];
        await page.goto(a.href, {timeout: 0});
        let filename = "items-" + i + ".png";
        await page.screenshot({
            path: filename,
            fullPage: true
        });
        page.close();
    }
    browser.close();
})();
```

2.数据持久化

- 根据爬取的规则和策略，把爬取到的数据储到数据库中
- 如果要兼容不同的来源，需要对数据需要格式化
- 为不同的数据建立索引方便检索

3. 数据订阅

- 用户可以按照自己的兴趣和需要进行订制内容

4. 分发

有新的数据的时候

- 可以使用邮件推送到订阅者
- 可以使用极光推送等推送服务
- 可以使用及时通信服务向客户端推送

参考

- axios (https://www.npmjs.com/package/axios)
- request (https://www.npmjs.com/package/request)
- puppeteer (https://github.com/GoogleChrome/puppeteer)
- puppeteer api (https://github.com/GoogleChrome/puppeteer/blob/v1.7.0/docs/api.md)
- showapi (https://www.showapi.com/api/view/184/4)