

网页数据爬虫——以知网为例


Yuxiang Wang

Hangzhou Dianzi University

目标：获取中国知网上各个学术期刊下的论文

提取相关信息并保存

手机版 English 帮助中心 登录 注册 我的CNKI NEW 充值中心 购买知网卡

 中国知网
www.cnki.net
中国知识基础设施工程

出版来源导航 ▼





来源名称 请输入检索词 出版来源检索 文献检索 >

出版来源导航 > 学科导航

学科导航

基础科学 > 工程科技 I 辑 > 工程科技 II 辑 > 农业科技 > 医药卫生科技 > 哲学与人文科学 > 社会科学 I 辑 > 社会科学 II 辑 > 信息技术 > 经济与管理科学 >

最近浏览 Recent Browse

 杭州电子科技大学学报(自然科学版)	 软件学报	 泥沙研究	 中国舰船研究
 机械工程	 船海工程	 农业开发与装备	 商用汽车新闻

项目

论文标题

分类号

论文摘要

论文作者

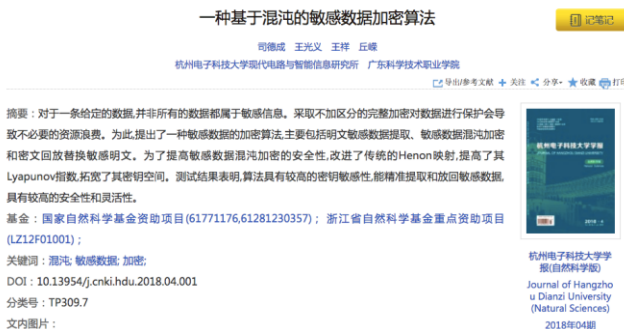
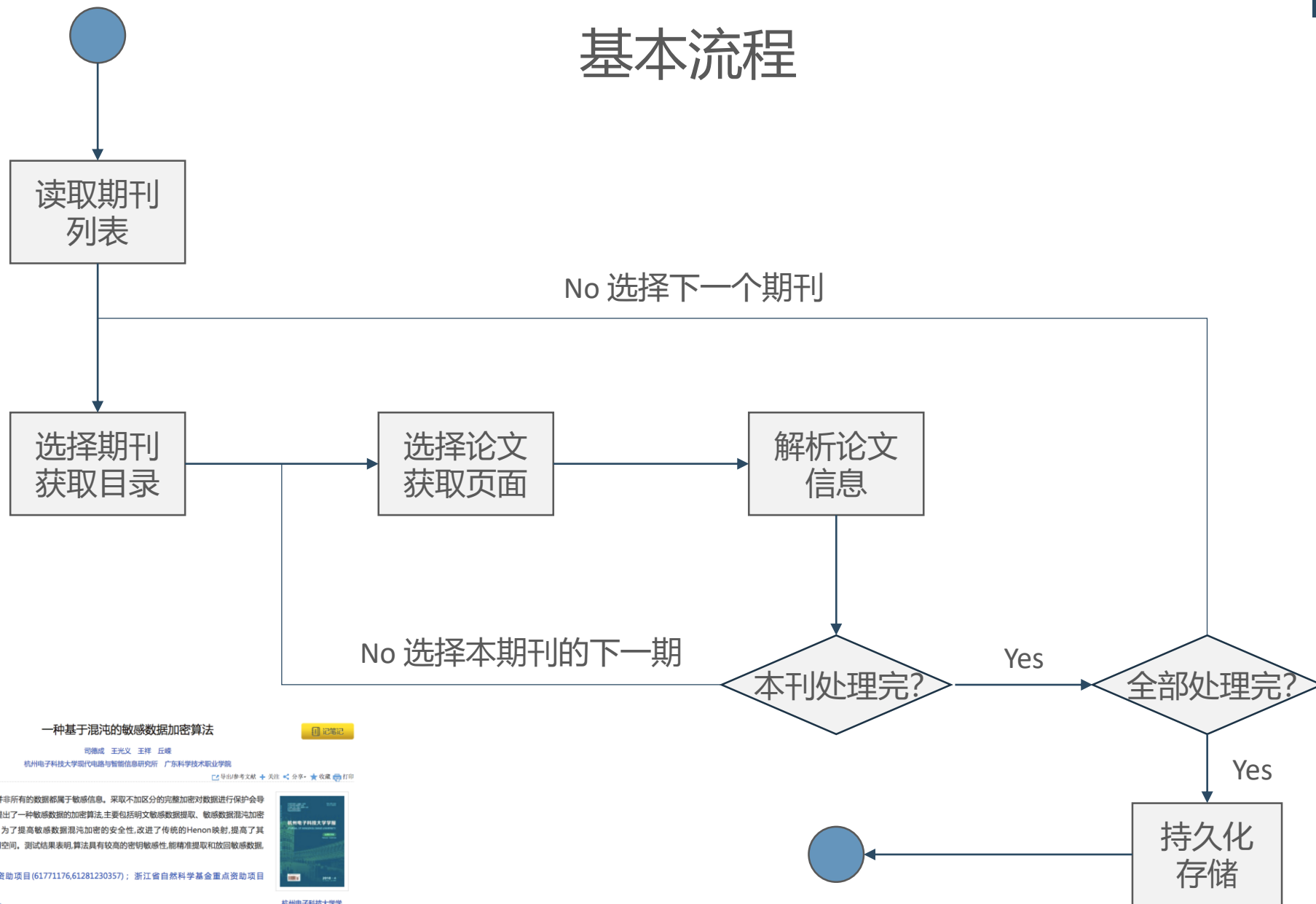
关键词

论文单位

基本流程



- 一种基于混沌的敏感数据加密算法
- 水下无线电能传输和信号接口系统设计和分析
- 一种基于连续度的自适应改进型碰撞树算法
- 等离子消融手术系统的全桥逆变输出电路设计
- Massive MIMO中一种低复杂度信道估计方法
- 无线密钥提取过程中改进的多比特量化算法
- 一种基于NLCS的斜视SAS成像处理算法
- 基于压缩感知的无线传感器网络定位算法研究



输入指定的期刊: HXDY, 年份: 2018, 期号: 4

```
if __name__ == '__main__':
    for dbcode, filename in get_paper_url_info(2018, 4, "HXDY", 0):
        print "===== "
        get_paper_info(dbcode, filename)
        print "===== \n"
        time.sleep(3) # 防止频繁访问造成IP被禁, 采用简单的访问一次等待一段时间

def get_paper_url_info(year, issue, pykm, paperId):
    """
    获取指定目录下的论文的dbcode和filename
    为获取论文做准备
    :param year: 期刊年份
    :param issue: 哪一期
    :param pykm: 期刊知网标识, 可以在resource下的journalnameCodes.txt的第三个字段中查到
    :param paperId: 第几页目录
    :return:
    """
    print "开始获取%s期刊 %s年 第%s期第%s页目录" % (pykm, year, issue, paperId)
    return PaperJournalReptile(PaperJournalReptile.url_create(year, issue, pykm, paperId)).parse_html()
```

根据指定参数拼接某期论文目录的URL

```
@staticmethod
def url_create(year, issue, pykm, paperId):
    """
    根据期刊指定的信息生成请求期刊目录的url
    :param year: 期刊年份
    :param issue: 哪一期
    :param pykm: 期刊知网标识, 可以在resource下的journalnameCodes.txt的第三个字段中查到
    :param paperId: 第几页目录
    :return:
    """
    url = "http://navi.cnki.net/knavi/JournalDetail/GetArticleList?year=%d&issue=%02d&pykm=%s&pageIdx=%d" \
        % (year, issue, pykm, paperId)
    return url
```

拼接URL的具体方法

实际获取的某一期论文目录的HTML如下

一种基于混沌的敏感数据加密算法

-
-
- - 新浪微博
 - 腾讯微博
 - 人人网
 - 开心网
 - 豆瓣网
 - 网易微博

司德成;王光义;王祥;丘嵘; 1-5+10

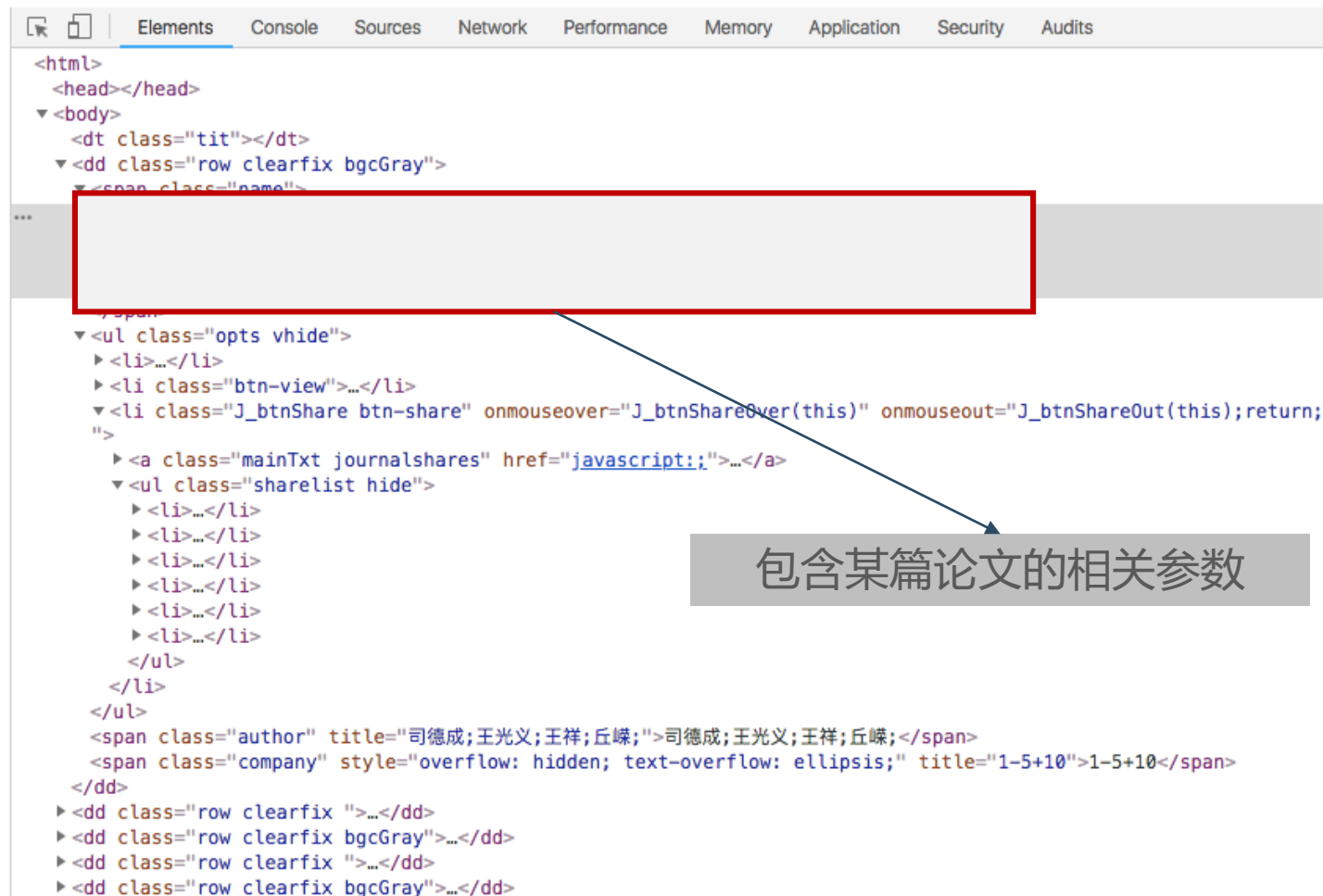
水下无线电能传输和信号接口系统设计和分析

-
-
- - 新浪微博
 - 腾讯微博
 - 人人网
 - 开心网
 - 豆瓣网
 - 网易微博

周世鹏;刘敬彪;史剑光; 6-10

一种基于连续度的自适应改进型碰撞树算法

-



```
<html>
<head></head>
<body>
  <dt class="tit"></dt>
  <dd class="row clearfix bgcGray">
    <span class="name">
      ...
    </span>
    <ul class="opts vhide">
      <li>...</li>
      <li class="btn-view">...</li>
      <li class="J_btnShare btn-share" onmouseover="J_btnShareOver(this)" onmouseout="J_btnShareOut(this);return;">
        <a class="mainTxt journalshares" href="javascript:;">...</a>
        <ul class="sharelist hide">
          <li>...</li>
          <li>...</li>
          <li>...</li>
          <li>...</li>
          <li>...</li>
          <li>...</li>
        </ul>
      </li>
    </ul>
    <span class="author" title="司德成;王光义;王祥;丘嵘;">司德成;王光义;王祥;丘嵘;</span>
    <span class="company" style="overflow: hidden; text-overflow: ellipsis;" title="1-5+10">1-5+10</span>
  </dd>
  <dd class="row clearfix ">...</dd>
  <dd class="row clearfix bgcGray">...</dd>
  <dd class="row clearfix ">...</dd>
  <dd class="row clearfix bgcGray">...</dd>
</body>
</html>
```


“一种基于混沌的敏感数据加密算法” 的论文页面URL

<http://kns.cnki.net/kcms/detail/detail.aspx?dbcode=CJFD&filename=HXDY201804001&dbname=CJFDLAST2018>

一种基于混沌的敏感数据加密算法

-
-
- - 新浪微博
 - 腾讯微博
 - 人人网
 - 开心网
 - 豆瓣网
 - 网易微博

司德成;王光义;王祥;丘嵘; 1-5+10

[水下无线电能传输和信号接口系统设计和分析](#)

-
-
- - 新浪微博
 - 腾讯微博
 - 人人网
 - 开心网
 - 豆瓣网
 - 网易微博

周世鹏;刘敬彪;史剑光; 6-10

[一种基于连续度的自适应改进型碰撞树算法](#)

-

```

Elements Console Sources Network Performance Memory Application Security Audits
<html>
<head></head>
<body>
  <dt class="tit"></dt>
  <dd class="row clearfix bgcGray">
    <span class="name">
      <a target="_blank" href="Common/RedirectPage?sfield=FN&dbCode=CJFD&filename=HXDY201804001&tableName=CJFDLAST2018&url=">
        一种基于混沌的敏感数据加密算法
      </a>
    </span>
    <ul class="opts vhide">
      <li>...</li>
      <li class="btn-view">...</li>
      <li class="J_btnShare btn-share" onmouseover="J_btnShareOver(this)" onmouseout="J_btnShareOut(this);return;">
        <a class="mainTxt journalshares" href="javascript:;">...</a>
        <ul class="sharelist hide">
          <li>...</li>
          <li>...</li>
          <li>...</li>
          <li>...</li>
          <li>...</li>
          <li>...</li>
        </ul>
      </li>
    </ul>
    <span class="author" title="司德成;王光义;王祥;丘嵘;">司德成;王光义;王祥;丘嵘;</span>
    <span class="company" style="overflow: hidden; text-overflow: ellipsis;" title="1-5+10">1-5+10</span>
  </dd>
  <dd class="row clearfix">...</dd>
  <dd class="row clearfix bgcGray">...</dd>
  <dd class="row clearfix">...</dd>
  <dd class="row clearfix bgcGray">...</dd>

```

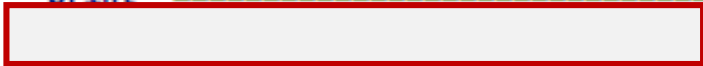
从目录URL中每个论文的href标签中可以解析出论文URL的参数

将目录HTML中所有的href标签提取出来，解析出对应的dbCode和filename

```
...<html> == $0
<head></head>
▼ <body>
  <dt class="tit"></dt>
  ▼ <dd class="row clearfix bgcGray">
    ▼ <span class="name">
      <a target="_blank" href="Common/RedirectPage?sfield=FN&dbCode=CJFD&filename=HXDY201804001&tableName=CJFDLAST2018&url=">
        一种基于混沌的敏感数据加密算法
      </a>
    </span>
    ▼ <ul class="opts vhide">
      ▶ <li>...</li>
      ▶ <li class="btn-view">...</li>
      ▼ <li class="J_btnShare btn-share" onmouseover="J_btnShareOver(this)" onmouseout="J_btnShareOut(thi">
        ▶ <a class="mainTxt journalshares" href="javascript:...">...</a>
        ▼ <ul class="sharelist hide">
          ▶ <li>...</li>
          ▶ <li>...</li>
          ▶ <li>...</li>
          ▶ <li>...</li>
          ▶ <li>...</li>
          ▶ <li>...</li>
        </ul>
      </li>
    </ul>
    <span class="author" title="司德成;王光义;王祥;丘嵘;">司德成;王光义;王祥;丘嵘;</span>
    <span class="company" style="overflow: hidden; text-overflow: ellipsis;" title="1-5+10">1-5+10</spa
  </dd>
  ...
```

```
def parse_html(self):
    """
    获取并解析论文期刊页面，得到dbcode和filename中的内容
    :return:
    """
    if self.tree is None or self.requestCode != 1:
        raise RuntimeError("%s页面解析失败" % self.url)
    if self.html.encode('utf-8').find("暂无目录信息") == -1: # 找到该目录信息
        paperUrlInfo = list()
        hrefs = self.get_elements_info('//span[@class="name"]/a[@target="_blank"]/@href')
        if hrefs:
            for href in hrefs:
                pattern = re.compile("dbCode=(.*?)&filename=(.*?)&") # 利用正则表达式提取出dt
                match = pattern.search(href)
                if match and len(match.groups()) == 2:
                    paperUrlInfo.append([match.group(1), match.group(2)])
            return paperUrlInfo
        else:
            print "%s: 该页面暂无目录信息" % self.url
            return None
```


拿到所有的dbCode和filename后，就可以访问每一个论文的面

```
if __name__ == '__main__':  
    for dbcode, filename in get_paper_url_info(2018, 4, "HXDY", 0):  
        print "===== "  
          
        print "===== \n"  
        time.sleep(3) # 防止频繁访问造成IP被禁，采用简单的访问一次等待一段时间
```

dbcode=CJFD&filename=HXDY201804001

```
def get_paper_info(dbcode, filename):  
    """  
    解析论文页面，将论文的信息放在一个字典内  
    :param dbcode:  
    :param filename:  
    :return:  
    """  
    return PaperReptile(PaperReptile.url_create(dbcode, filename)).parse_html()
```

<http://kns.cnki.net/kcms/detail/detail.aspx?dbcode=CJFD&filename=HXDY201804001&dbname=CJFDLAST2018>

获取每个论文的HTML后，对其进行解析以提取相关信息，并打印

```
<div class="wxTitle">
  <h2 class="title">一种基于混沌的敏感数据加密算法</h2>
  <a class="btn-note" target="_blank" href="http://x.cnki.net/search/common/te
dbcode=CJFD&tablename=CJFDLAST2018&filename=hxdy201804001&filesourcetype=1">...</a>
  <span>
    <a onclick="
      TurnPageToKnet('au','司德成','39627139');
    </span>
  <span>
    <a onclick="
      TurnPageToKnet('au','王光义','10978867');
    </span>
  <span>
    <a onclick="
      TurnPageToKnet('au','王祥','35428064');
    </span>
```

打印每篇论文所提取的信息

中国知网 cnki.net 期刊

请输入搜索内容 检索

一种基于混沌的敏感数据加密算法

司德成 王光义 王祥 丘嵘

杭州电子科技大学现代电路与智能信息研究所 广东科学技术职业学院

导出/参考文献 关注 分享 收藏 打印

摘要：对于一条给定的数据,并非所有的数据都属于敏感信息。采取不加区分的完整加密对数据进行保护会导致不必要的资源浪费。为此,提出了一种敏感数据的加密算法,主要包括明文敏感数据提取、敏感数据混沌加密和密文回放替换敏感明文。为了提高敏感数据混沌加密的安全性,改进了传统的Henon映射,提高了其Lyapunov指数,拓宽了其密钥空间。测试结果表明,算法具有较高的密钥敏感性,能精准提取和放回敏感数据,具有较高的安全性和灵活性。

基金：国家自然科学基金资助项目(61771176,61281230357)；浙江省自然科学基金重点资助项目(LZ12F01001)；

关键词：混沌；敏感数据；加密；

DOI：10.13954/j.cnki.hdu.2018.04.001

分类号：TP309.7

文内图片：

知识节点

- 基本信息
- 摘要
- 基金
- 关键词
- DOI
- 分类号
- 文内图片

知识网络

- 引文网络
- 关联作者
- 相似文献
- 读者推荐
- 相关基金文献

```
else:
    keywords = None
elif label == "catalog_ZTCLS":
    classification = catalog.xpath('./text()')
    classification = classification[0] if classification else None
self.paperDic['name'] = self.list_to_str(papername)
self.paperDic['authors'] = self.list_to_str(authors, ";")
self.paperDic['organization'] = self.list_to_str(organization, ";")
self.paperDic['keywords'] = keywords
self.paperDic['abstract'] = self.list_to_str(abstract)
self.paperDic['classification'] = classification
self.show_paper_dict()
return self.paperDic
```

杭州电子科技大学学报(自然科学版)

Journal of Hangzhou Dianzi University (Natural Sciences)

2018年04期

ISSN：1001-9146

存在哪些问题？

不支持多个期刊自动爬取

从期刊目录中选定几个期刊
进行自动爬取

不支持数据永久化存储

将解析好的论文信息
存储在文件中

没有异常处理

增加异常处理逻辑

没有代理IP支持

增加代理IP使用

没有多线程支持

增加多线程支持

增加异常处理逻辑

增加代理IP使用

修改完善之前的代码
实现这些功能

爬取计算机学报、软件学报
2010-2018年所有论文

爬取结果写入文件
为后续数据库写入做准备



1. 读取一个待爬期刊列表
2. 依次爬取

类似配置文件一样

论文标题: XXXXXX
论文作者: XXXXXX, XXXXXX, XXXXXX
论文单位: XXXXXX, XXXXXX, XXXXXX
论文关键词: XXXXXX, XXXXXX, XXXXXX
论文摘要: XXXXXX
论文分类号: XXXXXX
论文年份: XXX

论文标题: XXXXXX
论文作者: XXXXXX, XXXXXX, XXXXXX
论文单位: XXXXXX, XXXXXX, XXXXXX
论文关键词: XXXXXX, XXXXXX, XXXXXX
论文摘要: XXXXXX
论文分类号: XXXXXX
论文年份: XXX

.....

为避免单个文件过大, 可每满 N 论文就新开一个文件