# Design and Implementation of It Job Recruitment Data Based on Web Crawler

**RongFu Wang**

Guangdong University of Science and Technology, Dongguan, Guangdong, 523083, China

**Keywords:** Web crawler, It recruitment, Python, Analysis of data

**Abstract:** Crawler technology, as the most critical technology in data search technology, is also the core module in search engine, providing data source for search engine. In order to obtain a large amount of recruitment information existing on the Internet and reflect it locally, a web crawler system is designed to crawl this information and store it in categories. In this paper, Python web crawler technology is used to obtain job information of recruitment website and store it in database, XPath module is used to clean and crawl job data, and Struts2+hibernate is used to realize employment recommendation system. IT can be used as a reference for IT job recruitment.

## 1. Introduction

At present, all companies are facing a problem, and they don't know how to keep pace with the salary level of the market, which means that in the current market, the salary of IT staff is growing much faster than expected, and IT staff is still in short supply. Especially in the current competitive incentive of IT talent market, enterprises can not participate in the competition without principle, but should jointly maintain the market order, and attract talents by continuously strengthening the construction of corporate brands and enhancing corporate image; Build enterprise culture, and establish employee career development channels to retain talents [1].

Crawler can be understood as a program that can get the content of web page by itself. It can get the required information through the specified page and filter out other unnecessary information. Crawler is very important not only in the search field, but also in scientific research and in many large enterprises, crawler technology is used to obtain data. Python language is often used to write crawler programs, but compared with Java, Python has lower execution efficiency, and Java language has better support for multithreading [2]. A web crawler system based on Java language is designed for a recruitment website to crawl the relevant information of the platform users. At the same time, in order to analyze the data better, the system can make statistics on the obtained user data from multiple angles and present it in an intuitive chart form.

## 2. Structure and Basic Principle of Web Crawler

Web crawler is the key technology in search engine, which is mainly responsible for collecting all kinds of information on the Web. It usually uses hypertext links on the Web to access the corresponding web pages [3]. The web crawler will start crawling from a pre-established URL list, which may be extracted from historical access records, or some mainstream websites and web pages, and then crawl to other pages through these URLs by using HTTP and other protocols until all the URLs that meet the conditions are searched.

Web crawler technology is a common means to collect data and information. Because of its advantages of automation, expansibility and relatively simple development, it is widely used in data model analysis of all walks of life. The web crawler stores each link in the queue and calls it one by one [4]. When the link is used, it is automatically discarded. At the same time, the data crawled on each link is stored in the database for users to use.

Short path crawls deep resource data and reduces data collection cost [5]. The specific process is to determine a small number of URLs as child nodes, and store them in the URL queue to be crawled, then download the corresponding page data information from the corresponding URL site, and when these URLs are successfully crawled, put them into the crawled URL queue, and then

  DOI: 10.25236/icited.2021.052

extract new link addresses for page data mining. The process is shown in Figure 1.

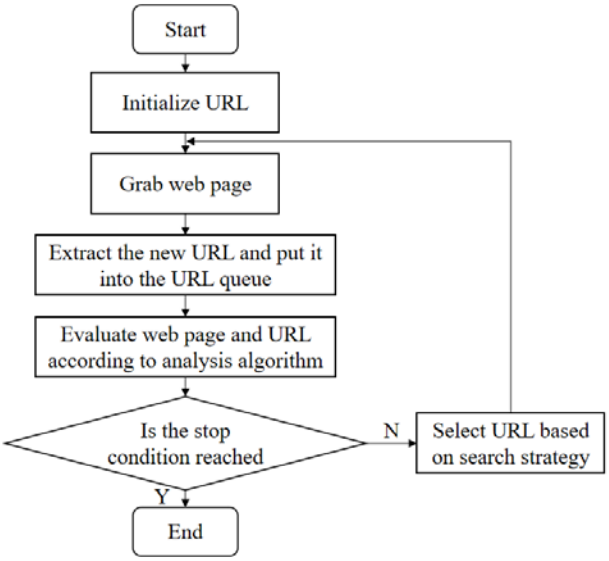


Fig.1 Flow Chart of Web Crawler

Because HTML has a relatively fixed file format, which is similar to the number of hyperlinks, titles and other information, the crawler can get it by analyzing the contents of HTML files [6]. All the information captured by the crawler will enter the index database, and the hyperlinks it contains will be used as the new starting URL. The crawler will repeat this process to collect information on the Web.

Under normal circumstances, the crawler in the search engine will visit the collected web pages regularly, detect the changes of the web pages, remove useless necrotic links, update the index database, and feed back the changes to the user's query results. So that users can see more concise results.

The crawling work of Web crawlers needs to be carried out according to certain strategies and algorithms. There are usually four traversal algorithms:

(1)Depth-First-Search

The basic principle is as follows: assume that M is the currently visited vertex, mark m as visited, select an undetected route (M, N) from M to N, and if it is detected that N vertex has been visited, go back and select another route from m that has not been crawled; If the vertex of N is not visited, visit N and mark it as visited, then search down from N, return to point M after detecting all routes distributed from N, and select another edge that has not been crawled to repeat the above process until all routes from M are detected.

(2)Breadth first search

Different from the depth-first algorithm, breadth-first-search will first crawl all the web links contained in the starting web page, then select one of the web links as the next starting page, and continue to crawl all the web links contained in this web page [7]. Compared with depth-first algorithm, this method can avoid endless loop caused by crawler searching deep all the time.

(3)Heuristic search algorithm

Originally originated from artificial intelligence, firstly, the value and relevance of web links to be visited are evaluated through certain domain knowledge and analysis strategies, and the distribution of information resources is speculated. Then, according to the established principles, the links with the highest value or relevance are selected for priority access, and the best route is found, with worthless nodes removed and high-value nodes retained. At present, this algorithm is mainly applicable to the search strategy of topic web crawler.

(4)Automatic classification search algorithm

It is also a search strategy of topic Web crawler, which makes the crawler have certain initiative, can reserve experience information, analyze information on the web, and analyze whether the page is the topic category that customers need for searching, so as to obtain the appropriate path for

searching [8].

## 3. Design and Implementation of It Job Recruitment Data

### 3.1 Analysis of Data

In this paper, according to the characteristics of recruitment information in recruitment websites, crawler cases are designed, and crawler tools are used to crawl recruitment data for later data analysis. First, use the browser to open the homepage of the recruitment website, enter the job keywords in the search box, such as “java Development Engineer”, and all the information related to the job will be listed, which is the data to be crawled.

The process of system development is mainly carried out in Windows environment, in which the development tools used in the system include Visio2013 for drawing system flow chart, E-R diagram and function module diagram, etc., MySQL 8.0.13 for building database, generating data table, etc., and realizing data storage. Python development tool chooses PyCharm to realize simple interaction and background business logic. The system runs under Windows environment. The data comes from the IT job recruitment website, and the main extracted and analyzed data are related contents of the recruitment website list page and details page.

Right-click the vacancy of the recruitment position, select the “Check” option, enter the source code debugging window, locate the corresponding label position, and keep a complete position information in the label. Data-index attribute indicates that each page can display up to 15 retrieved job information. Corresponding to company name, region, job title, working years, salary, company qualification, company scale, category, welfare, etc. After the above observation and analysis, these key values are data to be crawled.

In the selection of development library, the third-party library requests library in Python is an http library developed based on Python, which is highly encapsulated on the basis of Urllib library, which can not only read the returned data repeatedly, but also automatically determine the coding of the corresponding content, thus reducing a lot of work and being convenient to use.

### 3.2 Url Management

Set in Python environment is similar to set in other languages, and it is an unordered and non-repeating element set. Set basically supports relationship testing and eliminating duplicate elements. To prevent repeated crawling and cyclic crawling, especially when two URLs point to each other, the crawler will fall into an infinite loop. Use Python's set to manage URLs. The URL management process is shown in Figure 2.

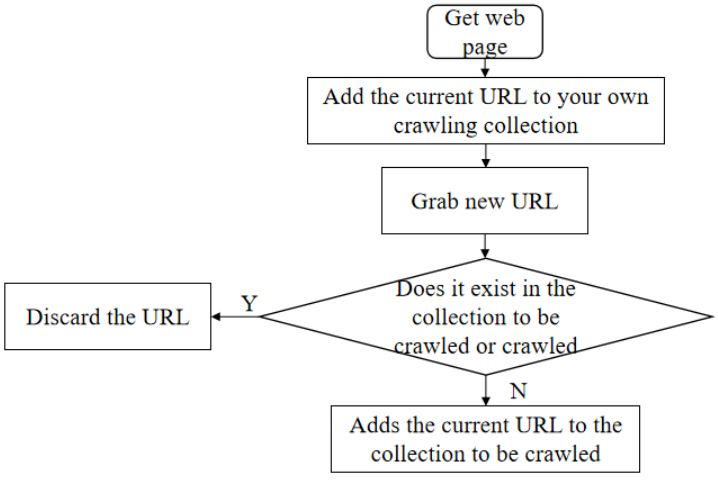


Fig.2 Url Management Process

### 3.3 Get User Data Details

Take out the user from the database table, get a user who has not been paid attention to, then

assemble the link address of the user information page, get all the detailed information of the user and store it in the data, and modify the user in the queue into the collected user. The main code implementation is as follows:

```
while (true) {
if (Const.FDI == 1) {
try {
Old one Info = old Mapper.get One Info();
String url=
“https://www.zhihu.com/people/” + one Info.get Name () +”/activities”;
old Mapper.up Info(one Info.get Id());
String zh_name = split[split.length - 2];
user Base Info Mapper.insert(user Base Info);
```

Through this system, the user data list details of social networking platform are obtained, and the obtained user details are presented in the form of a list.

In order to achieve the function of crawling web content, this crawler system adopts the open source package of HttpClient, which is modified on the basis of HttpClient according to the requirements of the project.

HttpClient provides a fully functional, convenient and up-to-date client multifunctional program toolkit suitable for HTTP protocol. At the same time, it also meets the current version of HTTP protocol in real time. HTTP protocol is a stateless protocol, which depends on request and response mode, and is usually based on TCP. Most Web development is based on HTTP protocol.

HttpClient supports requests by means of GET and POST, etc. GET requests ask for data from the server, while Post requests submit data to the server, and Get gets information instead of modifying information, which performs similar functions as database query. The specific workflow is shown in Figure 3.


Fig.3 Flow Chart of Web Page Acquisition in Crawler System

The web page crawling function in this program is realized by Com.spider.Crawler package, which includes the page content crawling class. The SingleCrawler is a single crawler, the CrawlerThread class is a multi-threaded crawler, and the crawler is the main crawling class. The implementation code is as follows:

```
package com.spider.crawler;
import java.io.IOException;
public class SingleCrawler {
public static String getContentOfPage(String url) throws IOEception{String res = null;
CloseableHttpClient httpclient = HttpClients. CreateDefault();
Try{
HttpGet httpget = new HttpGet(url);
// Execute HTTP request
CloseableHttpResponse response = httpclient.execute(httpget);
try{
res = EntityUtils.toString(entity,”GB2312);
entity.consumeContent();
}
} catch (IOExceptin e){
e.printStackTrace();
} finally {
response.close();
```

```
}
} finally {
httpclient.close();
}
return res;
}
```

### 3.4 Implementation of Data Analysis Module

The data content in the data analysis module comes from the content saved in the database in the data extraction module. The data analysis module queries the text data to be analyzed from the database, including the title and detailed description of the job seeker, and other data such as user name and passport. After Chinese word segmentation processing is performed on the text data, the value score of each data, i.e. each job seeker, is counted according to the preset scoring rules based on the segmentation results and other data.

As the extracted data are all related to IT job recruitment, it is necessary to add some game terms to the custom dictionary, and then segment the text on this basis, so as to improve the accuracy of the results. After editing the custom dictionary, save it as a. txt file. Note that the encoding format of the file here needs to be set to UTF-8.

After the word segmentation is finished, you still need to go to stop words. Stop words refers to words that often appear in the text, but whether the words themselves have exact meanings. Removing stop words can effectively help the system to improve the keyword density and make the extracted word segmentation results more concentrated and prominent. At first, read the stop words file and save it in the list through stop words = [line. strip () for line in open ('stop _ words. txt',' r', encoding =' utf-8'). read lines ()], then traverse the disabled word list and remove the stop words, and you can get the final word segmentation result.

After obtaining the final word segmentation result, it is necessary to make score statistics for this part of the content. In addition to Chinese word segmentation, Jieba library also has the function of part-of-speech tagging. after counting the number of word segmentation in accordance with bug related description dictionary and the number of nouns in word segmentation results, the basic scores of text parts are calculated according to the designed rules. You can choose the screening conditions when counting high-frequency vocabulary: 1. Climbing time; 2. Posting time; 3. Keywords; After selecting the filter conditions, enter the filter contents according to the selected filter conditions, where the time format requirement is: year-month-day; After the input of filtering content is completed, it is also necessary to input the words with the top N frequency; After all the above contents are entered, the statistical results will be given.

### 3.5 Implementation of It Job Recruitment Recommendation System

IT job recruitment recommendation system is developed based on Struts2+H ibernate framework, and is designed with service-oriented architecture. The functional structure of the system is shown in Figure 4.

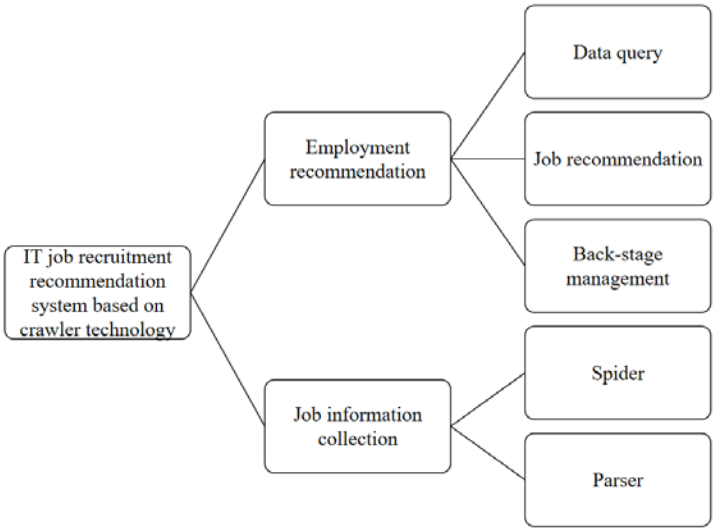


Fig.4 System Functional Structure Diagram

As we access the database based on the H ibernate framework, we need to write H ibernate tool classes, entity classes and corresponding mapping files, etc., besides importing Jar packages. It makes the system realize the basic operation of database efficiently and quickly.

### 3.6 Implementation of Url Deduplication Optimization Module

(1)URL deduplication

The solutions to the problem of URL duplication include storage based on hash algorithm, storage based on disk order, storage based on MD5 mapping and storage based on Bloom filter. Because the amount of data in this crawler system is not particularly large, the storage method based on hash algorithm is selected for URL storage. Through analyzing the recruitment information, we find that although the positions are different, the URL link addresses are regular, and they are all URLs spliced by ID numbers and other information. Therefore, the steps of URL duplication removal in this crawler system are to first process the URL, obtain the ID number in it, and then perform duplication removal by Hash algorithm, and the overall flow chart is shown in Figure 5.

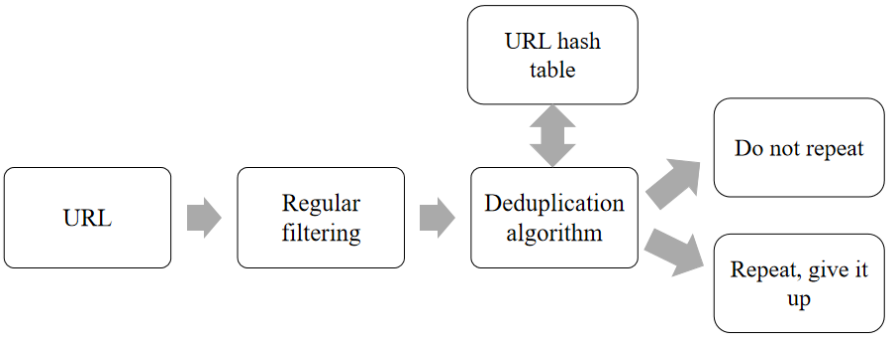


Fig.5 Url Deduplication Flow Chart

(2)URL filtering

In order to realize the filtering function of URL, we use regular expressions to filter, that is, filter all Html tags in the page through filtering rules. We put the filtering rules in the module, and establish the corresponding regular expressions for the content to be filtered, and the phrases in the expressions represent the features of the types to be filtered out. We need to extract http links from characters, so we need to set URL link regular expressions to complete the extraction and filtering of URL links. The filtering process using regular expressions is shown in Figure 6.

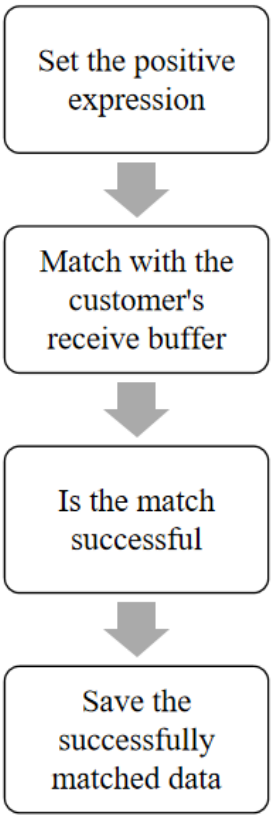


Fig.6 Process of Extracting Strings from Regular Expressions

(3)System optimization

Because the crawler is actually doing a lot of repetitive work, it makes the exchange frequency of memory and database high. In order to improve the working efficiency of this crawler system, we optimized its program. It mainly includes multithreading technology, JVM optimization, database optimization and offline script.

## 4. Result Analysis

In the crawler efficiency test, different thread numbers are used to test, and the test data are shown in Table 1. In order to save crawling time, the crawling search position page is set to 3 pages.

Table 1 Multi-Thread Test Table (Unit: s)

| Number of threads | Test 1(Java development engineer) | Test 2(Python) | Test 3 (front end designer) | mean time | Average improvement efficiency |
|---|---|---|---|---|---|
| 1 | 61 | 60 | 63 | 61.3 | |
| 2 | 45 | 44 | 40 | 43 | 30.25% |
| 3 | 33 | 36 | 35 | 34.7 | 17.63% |
| 4 | 31 | 34 | 32 | 32.3 | 0.65% |

It can be seen from the test results that when the number of threads changes from 1 to 2 and 3, the crawling speed is obviously improved. When the number of threads rises to 4, the system needs to allocate certain resources in order to maintain each thread; Considering factors such as network bandwidth, increasing threads has no obvious effect on crawler speed. Therefore, the use of multithreading technology needs to consider various factors such as machine performance and network bandwidth.

## 5. Conclusion

Aiming at the recruitment information in IT job recruitment websites, this paper provides a research and implementation of a web crawler to crawl the required recruitment information and

store it in categories. Through the cooperative efforts among teams, the functions of the crawler system are realized, and the required network data are successfully obtained.

## 6. Acknowledgment

Key Project of Natural Science in Guangdong University of Science and Technology (GKY-2019KYZD-7)

## References

[1] Li Y. (2018). Design and implementation of intelligent travel recommendation system based on internet of things. Ingenierie des Systemes d'Information, vol. 23, no. 5, pp. 159-173.

[2] Huang Guibin, Sun Liu, Huang Jialing, et al. (2018). Design and implementation of IT job recruitment recommendation system based on reptile technology market . Neijiang Science and Technology, vol.39, no. 01, pp. 62-64.

[3] Lee H E, rmakova T, Ververis V, et al. Detecting child sexual abuse material: A comprehensive survey. Forensic Science International Digital Investigation, 2020, 34:301022.

[4] Zhang B, Ye Y W, Shen X Z, et al. (2018). Design and implementation of levee project information management system based on WebGIS. Royal Society Open Science, vol. 5, no. 7, pp. 180625.

[5] Chang Fengjia, Li Zonghua, Wen Jing, et al. (2019). Design and implementation of recruitment data crawler based on Python . Software Guide, vol. 018, no. 012, pp. 130-133.

[6] Zhang Liangbin, Chai Hui, Wang Yuanming, et al. (2019). Continuous crawling and analysis of job data in recruitment websites based on distributed Docker cluster . Journal of Zhejiang Wanli University, no. 2, pp. 85-90.

[7] Tang Yihao. (2018). Application Practice of Web Crawler in Collecting Job Recruitment Data . Computer Knowledge and Technology, vol. 14, no. 28, pp. 14-15.

[8] Cui Zhaoxia, Liu Baolong. (2018). Design and Implementation of Web Data Crawler Based on Python . Digital Users, vol. 024, no. 016, pp. 10.