Innovative Application of Python in Data Crawling ---Chinese Version of Movie Recommendation Platform

Lishan Deng

International School, Beijing University of Posts and Telecommunications, Beijing 100876, China
denglishan@bupt.edu.cn

Abstract: The rapid development of computer hardware and Internet makes the information transmission speed and storage capacity of the network explode. The development of basic technology has led to the advance of many multimedia and rich text technologies that serve the public. In order to solve the problem that users and companies need to process and retrieve massive amounts of information, web crawler technology extracts the specified information from web pages and then analyzes it, providing a directional function of crawling information from large amounts of information. This article explains the web crawler that uses Python language to extract specific movie information from the website in detail and analyzes the detailed process of web crawler written in Python language.

1. Introduction
With the development of Internet technology, data has explosively increased in this era. According to the research report of IDC (International Data Corporation), the global data size will have reached 44ZB by 2020. Therefore Data Mining is an effective solution, which is widely used for gaining, storing, managing, and analyzing useful data in large databases [1].

This paper studies how to use data capture technology to get valid data in massive movie information data source, and then analyze users' movie preferences based on data that we captured to make accurate movie recommendations. This article introduces the data mining method of web crawler using Python language, and explain it from the specific introduction of web crawler, environment configuration, data acquisition, result analysis, and prospective application of data mining [2].

2. Literature Review
A web crawler is a program or script that automatically crawls specific information of web page in accordance with certain rules designed by the programme [3]. The crawler technology has been in development for a long time, and there are already many open source libraries and open source frameworks available. In this article, the open source framework of Scrapy, which is a web crawler framework written in Python that makes the design and work of crawlers simple and fast, will be used [4]. The Scrapy's architecture consists of eight components: Scrapy Engine, Scheduler, Downloader, Spiders, Pipeline, Downloader Middlewares, Spider Middlewares, and Scheduler Middlewares.

Widely-used web crawlers are mainly divided into the following types: general purpose web crawlers, focused crawlers, incremental web crawlers, and deep web crawlers. This article focused on the focused crawler. The focused crawler locates the crawled target webpage in a pre-defined topic-related page [5]. At this time, the bandwidth resources and server resources required for the...
crawler crawling can be greatly reduced.

Crawler applications exist in every aspect of life. For example, various industry data analysis, especially online applications tend to use web crawlers. Since the industry market forecast can be made according to the acquired data analysis [6]. At the same time, web crawlers can also monitor network comments and rob tickets on the internet.

3. Methodology

3.1. Environmental Preparation
Firstly, you need to install Python. Then you need to configure the environment variables. Scrapy is an application framework written to crawl website data and extract structural data. It can be applied to a series of programs including data mining, information processing or storing data. Installing pip and setting up related tools in the Python Package in advance are preferred. Installing Scrapy can be done using pip, typing “pip install Scrapy” at the command prompt. The structure of Scrapy is shown in Figure 1 below.

![Figure 1. Principle of Scrapy framework](image)

3.2. Structure of Program

3.2.1. Functional Module

3.2.1.1. Gaining of the Source Web

```python
def get_html(web_url):
    header = {
        "User-Agent": "Mozilla/5.0 (Windows; U; Windows NT 6.1; en-US) AppleWebKit/534.16 (KHTML, like Gecko) Chrome/10.0.648.133 Safari/534.16"
    }
    html = requests.get(url=web_url, headers=header).text  # response
    Soup = BeautifulSoup(html, "lxml")
    data = Soup.find("ol").find_all("li")  # filter and get the information needed
    return data
```

The first step is to get the header information of the webpage. The request.get() function returns response information based on the URL connection information. ".text" converts the response object to str type. Beautiful Soup is a Python library that can extract data from HTML or XML files. The find_all() function looks for all the "li" tags under the “ol” tag.
3.2.1.2. Information extraction and Processing

def get_info(all_move):
    f = open("E:/py_pro/result/douban2.txt", "a")

    for info in all_move:
        # rank
        nums = info.find('em')
        num = nums.get_text()

        # name
        names = info.find("span")
        name = names.get_text()

        For the ranking function, finding the "em" tag can get the rank of the target movie. For the movie
        name, finding the "span" tag can gain the movie name.

        # author
        charactors = info.find("p")
        character = charactors.get_text().replace("", "").replace("\n", "")
        charactor = character.replace("\xa0", "").replace("\xe0", "").replace("\xe6", "").replace("\xe0161", "").replace("\xe0f4", "").replace("\xe0f", "").replace("\xe207", "").replace("\xe5"

        To sort out author information, there are a lot of punctuation marks that interfere with data
        processing, so that needs to be replaced. After replacement, the information can be arranged more neat.

        # score
        scores = info.find_all("span", {"class": "rating_num"})
        score = scores[0].get_text()

        remarks = info.find_all("span", {"class": "inq"})
        if remarks:
            remark = remarks[0].get_text().replace("\xe2ef", "")
        else:
            remark = "No remarks"

        print(remarks)

    In the comments section, we need to make sure whether there are comments. Then the system can
    proceed with document processing or directly output "No remarks".

3.2.1.3. Score information analyzing with matrix

def Load_data(fh_train):
    fh_train = open(fh_train)
    trainSet = np.zeros((944, 1683))
    dict1 = {}

    for lines in fh_train:
        user, item, score, _ = lines.strip().split("\t")
        dict1.setdefault(user, {})
        dict1[user][item] = float(score)

        trainSet[int(user)][int(item)] = dict1[user][item]

    trainSet = np.delete(trainSet, 0, 1)
    trainSet = np.delete(trainSet, 0, 0)
fh_train.close()
return np.mat(trainSet)
fh_train = 'E:/py_pro/ml-100k/u1.base'
trainSet = Load_data(fh_train)
np.savetxt('trainSet.csv', trainSet, delimiter = ',')

3.3. Data processing method
In a large amount of web page information, the program needs to process the invalid information in order to find target information such as rank, movie name, director and comment. This article are able to filter the target site information that is crawled. The filter will get the movie information in the <li>tag of the web pages and then process it.

There are a large number of illegal information in the webpage that the web crawler cannot handle harmoniously, such as punctuation in the comment. Illegal information of this kind has been replaced in the project, so as to achieve unified information processing.

Since the data is continuous after the acquisition, it needs to be cut for later processing. In the program, the users, items and scores are divided into single units for data slicing. When processing the acquired data, the key and value in the database should be corresponding, and the value of the comment needs to be in accordance with the two-dimensional pair of user and item.

4. Discussion of Application and Function

4.1. Matrix Processing
During the data processing, this paper uses a matrix method for processing and analyzing. Firstly, this program creates a zero matrix and then import information such as users, movies, and comments into the matrix. This matrix explains the relationship between the user and the movie. Different scores represent different meanings. For example, a 5-point means that the movie is worthwhile to recommend. As shown in Figure 2 below.

|   |   |   |   |   |   |   |   |   |
|---|---|---|---|---|---|---|---|---|
| 5.00 | 5.00 | 4.00 | 3.00 | 5.00 | 0.00 | 4.00 | 1.00 | 5.00 |
| 4.00 | 0.00 | 0.00 | 0.00 | 6.00 | 0.00 | 0.00 | 0.00 | 6.00 |
| 0.00 | 0.00 | 0.00 | 0.00 | 6.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| 0.00 | 0.00 | 0.00 | 0.00 | 6.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| 0.00 | 0.00 | 0.00 | 0.00 | 6.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| 4.00 | 0.00 | 0.00 | 0.00 | 0.00 | 2.00 | 4.00 | 4.00 | 0.00 |
| 0.00 | 0.00 | 0.00 | 5.00 | 0.00 | 0.00 | 0.00 | 5.00 | 0.00 |
| 0.00 | 0.00 | 0.00 | 0.00 | 6.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 4.00 | 6.00 | 0.00 | 0.00 |
| 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 6.00 | 0.00 | 0.00 | 0.00 |
| 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 6.00 | 0.00 | 0.00 | 0.00 |
| 0.00 | 0.00 | 0.00 | 3.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 6.00 | 6.00 | 0.00 | 0.00 |

Figure 2. The matrix of Data Processing

4.2. Operation result

1. The Shawshank Redemption
Director: Frank Darabont Actor: Tim Robbins... 1994/ America/Criminal
Hope is a good thing, maybe the best of things, and no good thing ever dies.

9.6

Figure 3. The Format of Data storage
The project program in this paper will grab the movie's ranking, name, author, and comment from the
movie page. Then the program uses XPath to create a function that stores the acquired movie data in a .txt file in certain order. As shown in Figure 3

4.3. Analyze the problems encountered in the process, and how to deal with

In this paper, the problem of difficulty in identifying illegal character characters such as punctuation marks was encountered. The method of replacing illegal characters with recognizable and processed characters is used therefore.

A rating information as a value needs to be associated with two keys. By using this two dimensional key pair, the program can guarantee accurate and correct congruent relationship.

The verification process of some websites is a challenge such as verification code, dragging a picture and selecting pictures. And the innovative verification method is getting more and more difficult to deal with. The firewall of some website would limit the times of requests from the same IP. But usually the IP restriction is not designed to aim at the web crawler, it is used deal with DOS attacks. Agent IP which can request more, but it is still been limited. This problem cannot be solved absolutely.

Under some circumstance, the program would partly receive 403 error because the server does not respond to malicious attacks. Forged headers can be used here to which would help the program to camouflage as a web browser. Camouflage process includes finding the header from the website and copying it into the program.

5. Conclusion

Data mining plays a very important role in the case of explosive growth of data [7]. This paper begins with the introduction of the development of, and introduces the theory, flow path and application of data mining in detail. Meanwhile, this paper mainly talk about the web crawler as an important part of data mining. What’s more, this paper also describes in detail the development of web crawlers and their related architecture. This project uses the Python language to build a movie recommendation system based on movie ranking website information crawled by web crawler, and conducts testing and data analysis. This web crawler-based film recommendation project will greatly facilitate the audience who selecting movies and will work as a good example to lead the data mining technology servicing for our life. In the future, data mining will incorporate various data types such as graphics data, video image data, and sound data [8]. Data mining can excavate and analyze various data in life to provide more convenient support for all aspects of life.

Reference

[1] Wu, X., Zhu, X., Wu, G. Q., & Ding, W. (2014). Data mining with big data. IEEE transactions on knowledge and data engineering, 26(1), 97-107.
[2] Xiaofeng, M., & Xiang, C. (2013). Big data management: concepts, techniques and challenges [J]. Journal of computer research and development, I(98), 146-169.
[3] Fan, S., Lau, R. Y., & Zhao, J. L. (2015). Big data research.
[4] Zhour-un, Z. D. M. L. (2009). Survey of High-performance Web Crawler [J]. Computer Science, 8, 009.
[5] Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A. H. (2011). Big data: The next frontier for innovation, competition, and productivity.
[6] Data mining. (2018). Retrieved from https://en.wikipedia.org/wiki/Data_mining
[7] Liu Shilei. (2017). Coping strategies for anti-reptile sites. Computer Knowledge and Technology, 15, 010.
[8] Thelwall, M. (2001). A web crawler design for data mining. Journal of Information Science, 27(5), 319-325.