Volume: 10 Issue: 04 Pages: 3919-3924 (2019) ISSN: 0975-0290

Clothing Information Collection Based on Theme Web Crawler

School of Computer and Communication, Human Institute of Engineering Xiangtan 411104, China

Tang Zhi-hang

Email: zhtang@hnie.edu.cn

Li Jun

Email:642615662@qq.com

Zhou Yu-ying

Email:309856180@qq.com

----------**ABSTRACT**----------

With the rapid development of big data technology, many 'sleeping' data can be utilized, but the source of data is the key point. The previous methods of obtaining data can no longer meet the demand. This article uses python web crawler to down jacket of Alibaba International Station. Information (shell material, structure type, fill material, process information, and style information) is crawled and stored in the MongoDB database for data sources for apparel information analysis.

Key words: **Data mining; Python web crawler; Clothing information analysis; Down jacket**

Date of Submission: Sep 14, 2018 Date of Acceptance: Sep 28, 2018

I. Overview

With the advent of the era of big data [1], the maturity of data mining technology, more and more 'sleeping' data is fully utilized. This is especially true in the field of textiles and garments. In the case of down jackets, there are many styles of down jackets. However, due to the increased safety awareness of people, when choosing down jackets, not only the style of clothing, but also the main casing material and structure of down jackets. Some other influencing factors such as type, filler material,and process information. This makes the clothing details extra important and affects the volume of the clothes.

In order to better analyze the influence of the main outer casing material and structure type of clothing on the volume of clothes. A detailed introduction to web crawlers is given in this article. Implemented a theme web crawler with Alibaba International Station as an example，climb information on the classification of down apparel in Alibaba International Station, including shell material, structure type, filling material, process information and style information.

II. Analyze the design requirements of the clothing information collection system

2.1 Python web crawler

Web crawlers simply write programs to automatically retrieve data from web pages based on the rules of the web page. This program is called a crawler, which is a web crawler [2].Web crawlers can be implemented in many programming languages ， this article chose Python3[3],Python3 comes with a related crawler base library such as urllib，plus re regular expression base library、processing lxml library for XML and HTML、there is also a mongodb database to complete a theme crawler system.

2.2 Design requirements for the topic crawler system

When developing a web crawler system，First of all, we must analyze the system construction， design code specifications and functional modules for this system.In

Volume: 10 Issue: 04 Pages: 3919-3924 (2019) ISSN: 0975-0290

order to make the code easy to maintain and increase the importance of the code, you need to module the code to achieve their respective functions. Clothing information collection system is to obtain data by simulating request an item's detail page, and deposited into the mongodb.

III. Implementing clothing information collection system

3.1 Design of clothing information collection system

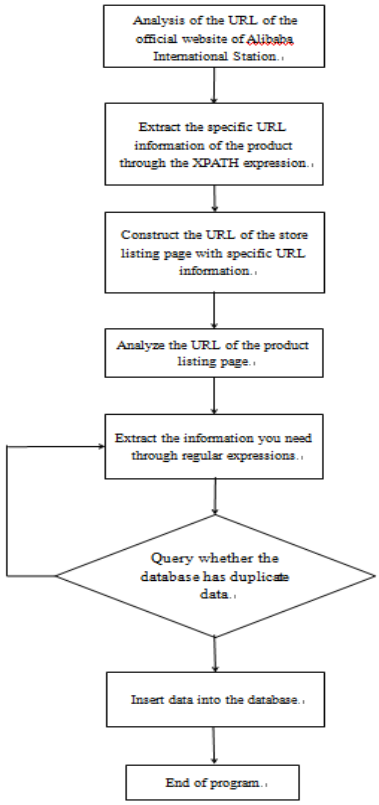

Figure 1Crawler crawling flowchart

The overall design idea of the clothing information collection system is as follows: This article takes Alibaba International Station as an example. First, we should clarify the URL page that finally obtains the data, then analyze the product details page, and draw important components according to the structure of the URL.From the back and forward speculation, we can first request the URL of the down jacket classified by Alibaba International Station, and then import the ertee in the lxml library to write the XPATH expression to extract the specific URL information of each product, and construct the product detail page through the specific URL information. The URL, finally requesting the product details page, analyzing the source code of the page, and then writing out the information needed for regular expression extraction, and depositing it into the MongoDB database. The specific flow chart is shown in Figure 1.

3.2 Page capture module

Page crawling is the first module that the entire system needs to execute, and it is also the key to the system. So before crawling the data, we need to carefully analyze the target page source code and post/get request, and use the developer mode to record the browser's User-Agent. Because most websites do not welcome crawlers, we need to access the page through the User-Agent emulation browser. If the number of visits is large, we can consider setting up multiple User-Agents, and accessing them in a round robin, effectively reducing the IP address of the blocked IP. Danger. For pages that need to be logged in to access, we can simulate the post request and the server conversation, and record the cookie value after login, use the returned session to maintain the session state with the web page, so that the python crawler can log in the page, and then The data is crawled.

The Alibaba International Station described in this article is relatively friendly and can be accessed by simply simulating the browser to capture the data on the page. But what we need is the data of the product details page. The URL of the details page has no obvious rules. At this time, the URLs of multiple detail pages can be put together for comparative analysis. It can be seen that the URL address is more complicated, and there is an irregular string of existence. By observing the source code of the product classification homepage, it is known that the product details page URL is in the source code, by first crawling the product listing page URL of the product classification page, and then crawling the product listing page data. When analyzing the webpage, by looking at the webpage source code and then looking up the keywords of the required data appearing in the webpage, an accurate positioning is performed. The source code of the webpage of the down jacket detail page of Alibaba.com is shown in Fig. 2.

Volume: 10 Issue: 04 Pages: 3919-3924 (2019) ISSN: 0975-0290

When you can't find the required data in the source code of the web page, you need to use the capture package to achieve the data capture. The commonly used packet capture tool is Fiddler. Fiddler can intercept, resend, edit, and transfer data packets sent and received by the network [8].It can also be used to detect network security. Open the Fiddler software and click on the web page to see the data conversation between the web page and the server. Use the search function to search for the keywords of the required information, and then accurately locate and then think about which way to extract the data. If you can't see the data you want, it means that the website has been encrypted. In this case, you can only use the developer mode of the web page to debug and analyze the encryption method of the encryption algorithm to find the encrypted variable. And then come up with a way to deal with it.

```
        <dl class="do-entry-item">
    <dt class="do-entry-item">
        <span class="attr-name J-attr-name" title="Brand Name">Brand Name:</span>
    </dt>
    <dd class="do-entry-item-val">
                        <div class="ellipsis" title="Down jacket">Down jacket</div>
                    </dd>
</dl>
        <dl class="do-entry-item">
    <dt class="do-entry-item">
        <span class="attr-name J-attr-name" title="Model Number">Model Number:</span>
    </dt>
    <dd class="do-entry-item-val">
                        <div class="ellipsis" title="DSC_0763">DSC_0763</div>
                    </dd>
</dl>
        <dl class="do-entry-item">
    <dt class="do-entry-item">
        <span class="attr-name J-attr-name" title="Fabric Type">Fabric Type:</span>
    </dt>
    <dd class="do-entry-item-val">
                        <div class="ellipsis" title="100% nylon">100% nylon</div>
                    </dd>
</dl>
```

Figure 2 Crawled page

3.3 Page processing module

After accessing and crawling all the data on the page, it is found that not all the data is useful to you. In this case, you need to further capture the data and eliminate the data without value. This is page processing. Page processing is usually done with regular expressions and XPATH expressions, and they have their own advantages and disadvantages for accurate data extraction. Regular expressions are the most commonly used. There are not many source code pages. When the required data is not wrapped, you can directly write regular expressions (.*?) to match the required information, which is simple and convenient. When the page source code is increased, and the structure is complicated and the data is distributed in multiple rows, using regular expressions to extract information requires writing complex expressions，this will only pay off. At this time, if you use the XPATH expression, you only need to accurately locate the useful information according to the label of the page source code. Here I analyze the source code of the search results page of Alibaba International Mall, and find that the specific information of the required product detail page is in the source code. The page source code is not complicated and can be directly extracted by regular expression. After extracting the product number, construct the product detail page. The URL, python simulates requesting the URL, and obtains the source code of the product details page. The analysis finds that the source code of the product details page is more complicated, and the required information is divided into multiple lines. At this time, the XPATH expression is selected to obtain the corresponding information. When developing a crawler system, only careful analysis of the source code of the web page is good at exploiting the advantages of the two expressions to make the entire crawler system more efficient. Choosing the right expression to extract the appropriate information can improve crawling efficiency.

There are many powerful libraries in crawler development, designed for different developers, such as the Beautiful Soup library, which can parse HTML and XML, handle non-standard HTML tags well, and then generate parsing trees. This saves developers time. There are also many extension libraries that can be used, and can be easily utilized as needed to easily write a fully functional crawler page processing module.

3.4 Data storage module

The amount of data crawled is very large, and storage in the database is the best choice. The database used in this article is MongoDB. Because of its high performance, easy deployment, and ease of use, MongoDB is very convenient to store data, making it the preferred database for crawler systems. In the process of crawling data to the database, there is an important problem to be solved, that is, the problem of data crawling repeatedly. There are many ways to get rid of duplicate data. This article operates the database from the code. The specific

Volume: 10 Issue: 04 Pages: 3919-3924 (2019) ISSN: 0975-0290

implementation process is to write the operation of the MongoDB database in the pipelines.py file in the scrapy framework. These methods have templates in the official documentation, just rewrite the method, and write self.dbin the process item method. User'].update ({'url_token': item ['url_token']}, {'$set': item}, True), this key code determines whether the user is stored by determining whether the user's unique identifier url_token exists. In the database, the returned data is item.

In addition to the MongoDB database used in this article, Python's built-in SQLite3 database, MySQL database, Redis database, etc. can be combined with Python to store the data obtained by crawling. These databases have their own advantages and disadvantages. You can take advantage of each other and choose the right database to store the data according to your needs.SQLite3 is a database that comes with Python. It can be used by direct import. It is simple and convenient to install the corresponding software. It is suitable for crawling crawler systems with small amount of data and fast crawling.Redis database is a must-use database for large crawler systems. When the amount of data is huge and the crawling speed is high, only Redis can do the job. Redis is a rich and powerful database and provides APIs in multiple languages. The main feature of Redis is a high-performance key-value database. This feature allows the Redis database to easily store json data on web pages.

IV. Analysis of clothing information collection results

After the clothing information collection system collects information into the MongoDB database, it can directly view the data in the database, or export the extracted json data into tabular data. View the data information that has been crawled through the database visualization tool Robo 3T, and then write the code mongo export -d alibaba -c user -o C:/Users/Administrator/Desktop/alibaba.csv--typecsv-f "Shell material, structure Type, fill material, process information and style information", alibaba represents the name of the database, user represents the name of the database table, and alibaba.csv represents the name of the exported file. In this way, the json data in mongodb can be exported to the excel table for easy viewing. For the data that is crawled, first observe whether the data is valuable, whether it is consistent with the result written by the crawler program, if it is inconsistent or has errors, it needs From the back to the speculation, find out which part of the problem has gone wrong, after the problem is found, modify, test, and finally a complete crawl, and then look at the overall data, the exported data is shown in Figure 3.

Shell Material:	Fabric Type:	Filling Material:	Technics:	Style:
100% nylon	100% nylon	90% goose down,10% feather	Plain Dyed	Jackets, Long down jacket
100%Nylon	Nylon	Down Feather	Printed	Jackets
100% Polyester	Woven	100% Polyester	Plain Dyed	coat
60%cotton 40%nylon	cotton & nylon	Down Feather	YARN DYED	Jackets
100% Polyester	100% nylon	Down Feather	Embroidered	Jackets
100% nylon	100% nylon+duck down	Down Feather	Printed	Jackets
Nylon	Worsted	Down Feather	Plain Dyed	Jackets
100% Polyester	Other	Down Feather	Embroidered	Jackets
POLYESTER / NYLON	Woolen	Down Feather	YARN DYED	Jackets, British, classical
100% Polyester	Canvas	100% Polyester	Printed	Jackets
100% Polyester	polyester	Down Feather	Plain Dyed	DOWN JACKET
Suede	Suede	100%Cotton/Fur	Plain Dyed	Jackets, winter warm jacket
100% Cotton	cotton	100% Polyester	Plain Dyed	Jackets
100% Polyester	polyester	Down Feather	Printed	Jackets
100% Polyester	dust coat fabric	100% Polyester	Plain Dyed	wholesale children latest dress style
100% Polyester	CUSTOM	100% Polyester	Embroidered	Jackets
100% Polyester	Worsted	Polyester, 100% Polyester	Other	Jackets
100% Nylon	420D Nylon	Down Feather	GARMENT DYED	Jackets
POLYESTER / NYLON	Print fabric	100% Polyester	Embroidered, Printed mens down jacket	Jackets

Figure 3 Crawled data

In Figure 3, you can clearly see that the required data information has been crawled. Observe the data to know that the crawled data is normal. Then you can use Python's pandas and numpy libraries to preprocess the data and clean up the existing ones. Interfere with the data, then use the clustering algorithm for data analysis, and finally use Python's matplotlib library to get the results of the visualization.

V. Notes on collecting data

This clothing information crawler system should pay attention to several aspects: (1) when a large amount of clothing data is needed, we need to set the delay function, and only after the delay can the data be crawled normally. Because the crawler speed is relatively fast, it will bring pressure to the other server. When the amount of data is large and there is no delay, the other server will find your ip access as the machine, so the server's firewall will block your ip, after that. You can't access the website for a while, which brings trouble to the crawler system. If you set the

Volume: 10 Issue: 04 Pages: 3919-3924 (2019) ISSN: 0975-0290

data collection delay function, the server is friendly, the server can't detect the web browsing or crawling program, so the whole crawler can be normal. run. (2) Establishment and use of IP proxy pool. Although the delay function is convenient and simple, the disadvantage is that the speed will be much slower. When you have fast crawling data requirements, you will not be able to use it. You can use the IP proxy pool. The same IP is usually used for a website for a short time. It will cause the IP to be blocked. However, the IP proxy pool refers to the use of multiple IP proxy loops to crawl data, which can solve the problem of IP being blocked. (3) Distributed crawler system. The crawler system usually crawls a lot of data, and naturally it takes a long time. The speed at which a computer can crawl data is limited. You can use a distributed crawler to solve it. The distributed crawler uses multiple computers. It's much more efficient to crawl data, and this distributed crawler can be implemented with the original scrapy framework [4] plus the Redis database. In addition to these aspects need to pay attention to, there are other details cannot be ignored, such as the crawler system that wants to do real-time crawling, we can run the crawler program on the remote server, so that the crawler program [5] can last The operation is continued without regard to interference from other factors such as power outages.

VI. Summary

This article uses a crawler framework Scrapy in Python to implement a clothing data information crawler system targeting Alibaba International Mall, which is efficient and easy to modify. It can crawl the product details of a certain category of Jingdong Mall into the MongoDB database, and then export it through the visualization tool, and further analyze the data. In the Internet age, there are many ways to develop crawler systems, and the languages developed are various. In order to facilitate the analysis of clothing data, Python is chosen as the development language, and a series of extension libraries such as urllib are used. The network library, the re regular expression library, the time time library, and the lxml parsing library, etc., jointly built the crawler system. Although the crawler system is fully functional, there are still many shortcomings, such as the need to manually change the code in the time processing, the data that is captured is not comprehensive, etc., and should be continuously improved in the process of crawling. In the development of the crawler system, due to the different request methods, source code and encryption level of the web page, many new problems will be encountered, but the essence is the same. If you know the principle, you can find a way to solve the problem. After crawling the data, you should also pay attention to the fact that the data can only be used for scientific research, and cannot be used for other purposes, otherwise disputes will arise. When developing a crawler system, the most important thing is to first know what kind of data information you want, and then think about how to get this data information. After selecting the website, analyze the source code of the webpage and use the appropriate method to crawl the data and deposit it into the database.

ACKNOWLEDGEMENTS:

Project supported by Provincial Natural Science Foundation of Hunan（2018JJ4047）

Reference

[1]Yang Wengang, Han Haitao. Archives Information Collection Based on Topic Web Crawler in Big Data Background [J]. LANT World, 2015, 478 (20):20-21.

[2]Jia Qiran. Design and Implementation of a Python-Specific Web Crawler [J]. Computer Knowledge and Technology, 2017, 13(12): 47-49.

[3]Jiang Shanbiao, Huang Kailin, Lu Yujiang, etc. Design and Implementation of Professional Web Crawler Based on Python [J].Enterprise Science and Technology Development, 2016 (8): 17-19.

[4]Li Daizhen, Xie Liyan, Qian Shenyi, et al. Design and Implementation of Distributed Reptilian System Based on Scapy [J].Journal of Hubei University for Nationalities (Natural Science Edition), 2017, 35 (3): 317-322.

Volume: 10 Issue: 04 Pages: 3919-3924 (2019) ISSN: 0975-0290

[5]Sun Bing.Design and Implementation of Python-based Multithread Network Crawler [J].Network Security Technology and Applications, 2018(4).

[6]Zhang Shuxin. Design and implementation of commodity information collection system based on web crawler[D]. Xiamen University, 2016.

[7]Zheng Zheng, Zhao Fei, Zhou Xin-hua.Research and Implementation of Enterprise Policy Information Collection Based on Theme Web Crawler[J].Computer Knowledge and Technology, 2017, 13 (5X): 49-51.

[8]Wang Bin,Zhang Yunwei,Liu Jian,et al.A design of web crawler for agricultural information subject[J].Anhui Agricultural Science,2009,37(20): 9699-9700.

[9]Liu Zhijie. Topic web crawler search strategy and topic identification method [D]. Wuhan Institute of Technology, 2017.

[10]Liu Jiancheng, Wu Baoguo, Chen Dong.Research and development of forest management knowledge acquisition system based on Web crawler[J].Journal of Zhejiang Agricultural and Forestry University, 2017, 34 (4): 743-750.

Biographies

Tang Zhihang was born in Hunan, China, in 1974. He earned the M.S. degrees in control theory and control engineering from zhejiang University of techonlogy, in 2003 and Ph.D. from donghua University China in 2009. At the same time ,he is a teacher in department of computer and communication, Hunan Institute of Engineering(Xiangtan, China) from 2003.Chaired the 49th China Postdoctoral Science Foundation grant, presided over science and technology projects in Hunan Province in 2010, presided over the Education Department of Hunan Province in 2010 Outstanding Youth Project, as the first author more than 30 papers were published.His current research interests include intelligent decision and knowledge management