Computer Life (CPL) ISSN: 1819-4818
DELIVERING QUALITY SCIENCE TO THE WORLD

Research on Character Portrait Technology Based on Reptile Technology

Huawei Mei[1, a], Di Liu[1, b]

Department of computer science, North China Electric Power University, Baoding, China

[a]3094998@qq.com, [b]1246478371@qq.com

Abstract: In the era of mobile Internet, the use of Internet products by users has soared. And because of the rise of online social networking, users will leave a lot of opinions on the network, such as like comments. Collecting and analyzing this information can extract the user's behavior habits and distinguish the user groups. The crawler can obtain information in a batch and cross-platform manner, which can provide a more comprehensive analysis of the platform user groups. This article is based on the scrapy web crawler, which crawls different users according to the comments of the web forum and the followers. The tensorflow framework is used to implement the K-means method to cluster the collected users, and finally to realize user images and even predict user behavior. Mainly achieve the following work:1) Set up the running environment of the scrapy framework and the tensorflow framework.2) Use scrapy to crawl website user data.3) Use the tensorflow framework to build a model for cluster analysis.

Keywords: Scrapy, web crawler, cluster analysis, user portrait.

---

1. INTRODUCTION

1.1 Subject background

This is an era of rapid development of computer technology. With the rapid development of computer technology, network technology has changed and affected people's lives and working methods. Obviously, each of us lives with the development of network technology. The change. Not inferior to the impact on modern society, different Internet technologies have an indelible impact on people's way of thinking.

The first is the rise of online social networking, from professional chat tools to a cottage web game, from well-known e-commerce platforms to the user interface of an electronic product, major businesses are trying to move real social networks into network products. This greatly expands people's social networks, and it also allows people to put a lot of time and energy into the network.

Then there is the promotion of the public's right to speak. Since the Internet became popular, people have had the opportunity to make their voices in different places. Post it, Weibo, knowing that there are all kinds of forums, there are all kinds of people to explain their opinions all the time, people with different opinions may argue for them, and people with the same viewpoints There may be some

differences. The collision between these views has inspired people to express their desires. Over time, more and more people tend to make their own voices about a certain problem.

Finally, the generation of the recommended algorithm. At first, in order to retain more users, Internet products consciously recommend their favorite content to users. This allows users with the same preferences to have more opportunities to focus on each other and eventually form small circles with the same hobbies. However, the recommendation algorithm recommends people with the same opinion and also shields those with the opposite view. Under this kind of shielding, people can easily create an illusion that most people agree with what they are doing and agree with what they think. This makes it difficult for people who elaborate on the Internet to question their own ideas, and makes people's views on different issues on the Internet more vivid.

1.2 Purpose and significance

1.2.1 Research purposes

Today, there are words on the Internet that hold different periods, different groups, and different opinions. Different perspectives are based on different user groups' thinking patterns at different times. This topic hopes to collect, classify, and study the number of people who have different opinions on different platforms, and how some people on a certain platform have different opinions on different issues. By analyzing the distribution of these users, the user groups on each platform are inferred.

After implementing cluster analysis on users of different platforms, we try to predict the general view of an event on different platforms and the possible views of a certain person on a certain problem.

1.2.2 Significance

If we can distinguish which type of group a user belongs to on our platform, we can make more precise advertising. For the whole, if we can know that a certain group has a particularly strong demand for a certain problem, we may be able to extract an entrepreneurial basis.

In addition, when a popular hot spot appears, we can judge the level of attention of different hotspots among different groups, and determine which kind of people are more inclined to that kind of viewpoint in real life. This will be of great help to other studies in other social sciences.

1.3 Research status

In recent years, with the rise of technologies such as machine learning, there have been many researches on user portrait technology through cluster analysis and even neural networks. Some people set out to express the user's speech as a word vector, which is analyzed by the user's emotional color. The other part studies the social network between users, and calculates the similarity of the user to achieve the study of user portraits.

2. RELATED DEVELOPMENT TOOLS

2.1 Introduction to the Scrapy Framework

Scrapy is an open source web crawler framework maintained by Scrapinghub, written in the popular Python language. At the beginning of the design, he was used to automatically obtain information on the Internet, download pictures on the network, and replace the real person to complete a lot of repetitive work.

Scrapy has a built-in Twisted asynchronous architecture. This is a mature Python asynchronous framework. On the one hand, it allows Scrapy to take full advantage of the latency of network IO. On

the other hand, the interface supported by the built-in framework itself allows Scrapy to fulfill various requirements in a more flexible way.

2.2 Introduction to the Tensorflow framework

The new generation of deep learning system TensorFlow, developed by Google, is based on DistBelief. First, the Tensor actually refers to an N-dimensional array, while the Flow Flow refers to a calculation based on a data flow graph. As the name suggests, it is a process of moving tensors. Flow from one stream in TensorFlow to another stream. TensorFlow is a system that provides complex data structures for analysis and processing of invisible brain networks.

TensorFlow has a high rating for audio processing, graphics ratings, natural language advice and processing systems. As a deep learning framework, it supports mobile platforms such as Linux, Windows and Mac platforms, and even mobile platforms. At the same time, TensorFlow provides an extremely rich API for deep learning. In most in-depth learning frameworks, the API provided by tensorflow is very comprehensive. The construction of basic units such as convolution kernel networks and neural networks, visualization tools can be easily implemented.

2.3 Introduction to MongoDB

2.3.1 Non-relational database

Modern computing systems generate an amount of data on the network at all times. A significant portion of this data is stored by relational database management systems (RDBMSs), and its rigorous and mature mathematical theory foundation has made application programming and data modeling simpler. However, with the rise of the Internet and the wave of informationization, traditional RDBMS began to find many problems in some industries. Web-based products have an increasing demand for database storage, and many database collections have to be used to solve the problem. At the same time, in today's big data, a lot of data will be read in and often read, but rarely modified, and the RDBMS will work inefficiently. In addition, the uncertainty caused by the business in the Internet era has led to frequent changes in the storage mode of the database. The rigid storage mode increases the complexity and difficulty of operation and maintenance.

On this basis, non-relational databases are becoming more and more popular. The non-relational database proposes another new concept, for example, stored in a key-value pair, its structure is not fixed, allowing each tuple to have different fields, and each tuple can be added according to its own needs. Adding a unique key-value pair for an element makes the database not limited to a fixed structure, nor does it affect other data because of the special circumstances of some data, thus reducing the time and space overhead. In this way, users can add certain fields according to their respective needs. In order to obtain information about different users, it is not necessary to perform an associative query on multiple tables as in the previous relational database. You only need to take the corresponding value according to the id to complete the query operation. However, non-relational databases cannot provide complex queries like SQL because they have relatively few connections between data. And it is difficult to reflect the integrity of the programming. Therefore, he is only suitable for storing some relatively simple data. If the program needs to query complex data, then it is recommended to use a traditional SQL database.

2.3.2 Introduction to MongoDB

MongoDB is a distributed open source database system. Support for high load conditions, and developers can dynamically add more nodes to ensure server performance. MongoDB brings a new way of storing for current web applications. MongoDB stores data as a document, and the data structure consists of a dictionary. Its documentation is similar to a JSON object in a web request. It can be said that the appearance of MongoDB seems to be a document full of json, which makes it a relatively simple operation. If developers need to make database mirroring, MongoDB's built-in support allows developers to do this over the network, making MongoDB more scalable than traditional databases. If you need to introduce new requirements or need more resources to expand your business in later development, MongoDB can easily meet these needs by adding new nodes. This means that it can be distributed on other nodes in the computer network. It is written in C++ which makes it possible to call directly to the underlying device. Mongo can also be queried using rich expressions. At the same time MongoDB also supports the current mainstream programming language.

2.4 K-means算法

In the k-means algorithm, k represents the number of clusters, and means represents the mean of the data objects in the cluster (ie, the description of the cluster center). Therefore, the k-means algorithm is also called the k-average algorithm. The k-means algorithm is a partition-dependent clustering algorithm. It uses distance as a measure of the measure of similarity between data objects. The smaller the distance between data objects, the higher their similarity, and the more likely they are in the same cluster. There are many ways to calculate the distance between data objects. In general, the k-means algorithm uses Euclidean distance to measure the distance between data objects.

2.5 Web Crawler

2.5.1 Web crawler introduction

A web crawler is a script or program that automatically crawls information on the web according to specific rules. In addition, he is also known as ants, automatic indexing, simulation programs or worms. The crawler can automatically crawl the search engine's network resources. After crawling, users can search for the resources needed on the network through search engines. The search engine is a very large and complex algorithmic system. The accuracy and efficiency of the search is very high on the system. In addition to being used for searching, reptiles can do other work. For example, Internet projects primarily analyze data and obtain value data by crawling related data.

2.5.2 Html page

HTML is a professional markup language for front-end web design. People are used to building webpage skeletons with html, specifying webpage styles with css, and then adding events to webpages using javaScrapt. Due to its low development cost, the language is widely used in various situations on the Internet. From the mobile app to the web application, wherever the browser can run, there is no shortage of html.

A large part of the content is crawled from the html page when the crawler crawls. The html page containing the data item is also one of the main contents of the http request. Html is called Hypertext Markup Language because its text contains hyperlink points. The crawler parses the web page, essentially extracting the specified elements on the html page and then splitting them.

2.5.3 Session

In computers, especially in network-related applications, Session is also known as session control. The Session object is used to store the configuration information and properties required for a particular user session. Thus, when a user jumps between pages of an application, the variables stored in the Session object are not lost and will remain in the entire user session. When a user requests a web page from an application, the server automatically creates a session object if the user has not yet created a session. When the session is abandoned or expires, the session will be terminated by the server.

The specific operation is as follows: After the HTTP request page, if the open session is used, the server will read whether the PHPSESSID in the cookie exists. If it does not exist, the server will generate a new session_id, which is first stored in the PHPSESSID in the cookie. And then generate a sess_ prefix file. When writing $_SESSION, the server serializes the write data into the sess_ file. When reading a session variable, it will first read the PHPSESSID in the cookie, get the session_id, and then go to the sess_sessionid file to get the corresponding data. Since the default PHPSESSID is a temporary session, it will disappear after the browser is closed, so when we revisit it, we will generate the session_id and sess_ files.

2.5.4 HTTP request header

When a browser requests a web page, it will implicitly send a portion of the information to the server. This information cannot be read directly but is handled as a request header by the server. Where Accept specifies the type of processing that the request needs to process. Authorization is equivalent to a key issued by the server for the browser to indicate its identity. User-agent is used to identify the type of browser. For example, a large website will send different web page data for the mobile terminal and the web terminal. Some older websites will also return different pages for different browser versions.

Similarly, there are also cookies, Content-Length, Connection, and other fields in the request header.

2.5.5 robot.txt

Robots.txt is an ASCII-encoded text file stored in the root directory of the website. It usually tells web search engine robots (also known as web spiders). What content on this site is a robot that search engines should not use. What is obtained and what can be obtained by the robot. Since the URLs in some systems are case sensitive, the file name of robots.txt should be lowercase. The robots.txt should be placed in the root of the website. If you want to define the behavior of search engine robots when accessing subdirectories individually, you can merge custom settings into robots.txt in the root directory or use robot metadata (metadata, also known as metadata). In this file, developers can declare parts of the site that they don't want to be accessed by crawlers so that search engines can access and include some or all of the site's content, or search engines can only be specified by robots.txt. The specified content. The first file that a search engine crawls on a website visit is robots.txt.

In addition, the main role of robots.txt is to ensure network security and website privacy. Regular crawlers follow the robots.txt protocol. With the plain text file robots.txt created in the root directory, the website can declare which pages the crawler does not want to crawl and which pages to include. Each website can control whether the website is willing to be included in a search engine such as

Baidu, or specify that the search engine only contains the specified content. When a search engine accesses a site, it first checks for the presence of robots.txt in the root of the site. If the file does not exist, the crawler will crawl along the link. If it exists, the crawler will follow the file. The content determines the scope of access.

2.6 Development tools

2.6.1 Pycharm

Developed by Jetbrains, Pycharm is an excellent Python IDE. While writing the code, pycharm intelligently fills and corrects the errors based on the code being hit. This also means that pycharm consumes more running memory than other programming software. In the support of the pair, pycharm can also easily obtain different versions of third-party libraries from the network. At this time, the researchers saved the configuration time and avoided the interference between different Python versions. In addition, pycharm also supports code highlighting, which greatly improves the readability of python code.

Currently updated to Pycharm 3, two versions are released: Professional Edition and Free Community Edition. The Professional Edition is a paid version that offers more advanced extensions, while Free Community Edition is a free version with no trial time limit. If the user is not very necessary to use the paid version of the advanced features, the free version is already qualified for most of the work.

2.6.2 Mongo Manangement Studio

Mongo Management Studio is a free MongoDB GUI tool for database management. It is a lightweight, clear interface that makes it easy to develop MongoDB-based projects. It was developed using nodeJs, Electron Framework, MongoDB and AngularJs. The free version is only available for Windows; businesses and individuals are available for Linux, Windows and MacOS. The Enterprise Edition (Web Server) supports the MongoDB web interface HTTP GUI, so Mongo Management Studio supports installation on the primary server and can then be accessed on any system that uses the browser locally or remotely.

2.6.3 Python

Python is a very popular programming language, actually created by Gido Van Rossum, which was released as early as 1991. Python quickly gained the favor of researchers with its grammar of introduction and good readability. And the researchers began to spontaneously write a third-party library for researchers that is convenient for researchers. With such a virtuous circle, Python has become a household name in the field of computer science.

Like dynamic typing programming languages, Python has a dynamic typing system and a memory reclamation mechanism that automatically manages memory. However, the original intention of Python is to facilitate writing and easy to expand, so in special cases, it is not necessarily comparable to C++ and other languages for the utilization of computer resources.

The Python interpreter itself runs on almost all operating systems. An Python interpreter, CPython, is written in C and is a community-driven freeware managed by the Python Software Foundation.

Python has a good performance in different areas. In the field of data analysis, Python has a well-developed data analysis ecosystem. Even the discussion posts of many researchers suggest that the research tools should be migrated to the Python ecosystem earlier. Numpy & Scipy, Pandas, and matplotlib are the trojans for Python's most basic data analysis. In the field of data collection, Python's

standard library urllib can collect some simple data on the network. The beautifulsoup is a very good html parsing tool. The Python crawler framework pyspider, scrapy, etc. are also very popular. In the field of web development, python owner django, flask and other good development framework.

3. INTERNET CRAWLER CONSTRUCTION BASED ON SCRAPY FRAMEWORK

3.1 Design requirements and analysis of web crawlers

3.1.1 Simulate the request headers required by the website

Take a question forum as an example. Checking the content of the website reveals that Response Headers and Request Headers are response headers and request headers, respectively. The Request Headers are the information that the browser carries when sending a request to the website. Similar to the ID card, the opposite website determines whether to accept your request by judging the information. The Response Headers are the information that the web page carries when it responds.

The most important of the many entries in the header is User-Agent. As the name suggests, it refers to the source of requests from website visitors. Different browsers correspond to different user-agents, and the background will judge what to return based on the content of the request. Web version. For example, mobile and desktop versions, and compatibility with different browsers.

There is a cookie in the request headers, which means that the browser will carry the cookie attribute when you request the web page. The website background will verify the user's login request in this way. At the same time, cookies also save a lot of other information that the website needs, such as temporary user habits.

3.1.2 Crawl target analysis

The source code returned by the web page is shown in Figure 3-1. The background will return the corresponding content. The part of the target website needs to be parsed to return the html request. Here, the corresponding element or attribute is extracted according to the location of the html tag. The specific operation is: first open the browser, through the element check to determine the parent tag of the element tag until the root tag. The regular expression is used to extract the data needed for the extraction from the character stream in the returned page.

The other part of the information is requested by the webpage to the background, and then the content of the webpage is dynamically loaded by javascript. This part needs to analyze the request of the webpage and ask for json data from the background. After getting the json data passed in the background, it is parsed by python.

3.2 Reptile design

3.2.1 Primary crawler design

The main crawler is the core part of the data acquisition framework. It is first spliced according to the http request format, and then the returned json is parsed. Extract the elements you need.

3.2.2 data element item design

Item is the basic unit for storing data. That is, the basic unit of data after crawling.

3.2.3 Database connection part design

Data processing is implemented by the pipline part and needs to be stored in the specified mongodb table.

Fig. 1 returned page source code

4. CONCLUSION

The user can be differentiated to a certain extent by the user's behavioral characteristics, but it is difficult to construct an ideal model by simply using the K-means algorithm. An improved algorithm for user clustering can be considered to make the neighbor set more similar to the target user. The final recommendation system predicts the user's project score mainly based on the user's nearest neighbor set, so the accuracy of the nearest neighbor set is also important for the quality of prediction. If the user is classified by the K-means algorithm, the nearest neighbor set of the user is found according to the similarity of the user from the cluster of the target user. The user's nearest neighbor set can be made closer to the target user, and the trend of interest of the predicted target user is more accurate.

ACKNOWLEDGEMENTS

During this period of study, I would first like to thank Mei Huawei. During the learning process, Teacher Mei patiently guides me. After work, I often help me to comment on the paper. I sincerely thank him for his guidance.Secondly, I would also like to thank my friends around me for helping me. I also thank the brothers who played basketball with me. When I am in trouble, I can take the initiative to ask me to go out to play and relieve my mood. Finally, I would like to thank my parents for their support of my education and my study and study.

REFERENCES

[1] Yun Yang. Design and Implementation of Scrapy-based Web Crawler[J]. Computer Programming Skills & Maintenance, 2018, (9): 19-21, 58.

[2] Liu Yu, Zheng Chenghuan. Research on Deep Web Crawler Based on Scrapy[J]. Software, 2017, 38(7).

[3] Tao Xinghai. Implementation of Web crawler simulation login website based on Scrapy framework [J]. Digital users, 2017, (6).
[4] Wang Lei, Liu Xiaodan. Design and Implementation of Scrapy-based Web Crawler System Framework[J]. Microcomputer Applications, 2019, 35(4): 48-50.
[5] Wang Fang, Zhang Rui, Gong Hairui. Design and Implementation of Distributed Crawler Based on Scrapy Framework[J]. Information Technology, 2019, (3).
[6] Peng Jiben, Wu Lin, Chen Xian, et al. Design and implementation of network negative emotion mining system based on crawler technology[J]. Computer Applications and Software, 2016, (10).
[7] Lü Weiping, Zhang Xiaomei. Application of Clustering Analysis Based on SPSS[J]. Fujian Computer, 2013, 29(9): 20-23.
[8] Wu Jianlan. Research on Sina Weibo crawler based on Python[J]. Wireless Interconnect Technology, 2015, (6): 93-94.
[9] Zhou Zhonghua, Zhang Huiran, Xie Jiang. Sina Weibo data crawler based on Python[J]. Journal of Computer Applications, 2014, 34(11).
[10] Chen Lin, Ren Fang. Design of Sina Weibo Data Reptile Program Based on Python[J]. Information Systems Engineering, 2016, (9).
[11] Wang Jinfeng, Peng Yu, Wang Ming, Zhong Sheng, Zhao Xuehui. Sina Weibo Data Grab Technology Based on Web Crawler[J]. SME Management and Technology, 2019, (1): 162-163.
[12] Chen Wei, Lan Dingdong, Ke Wende, Li Shujun, Deng Wentian. Research and Implementation of Sina Weibo Reptile Based on Java[J]. Computer Technology and Development, 2017,27(9):191-196.