An indirect network acquisition system based on Weibo user data

Zhizhuang Li*, Yuhao Zhang, Xueying Li, Chunlan Zhu
Artificial Intelligence and Student Development Research Laboratory, Long-Spring Education Group, Kunming 650500, Yunnan, China

ABSTRACT

The theory of the six degrees of separation states that there are no more than six people between someone and any stranger, meaning that someone can meet any stranger through no more than six intermediaries. As one of the most popular social media applications in China, Weibo and its user data can be employed as the testing tool for the practical significance the theory holds. In this paper, an indirect network acquisition system containing two subsystems: Weibo user data acquisition system and user relationship analysis system has been established by Web crawler, and the system's availability and effectiveness have been tested by system service test, which shows the possibility of the Indirect Network Acquisition System.

Keywords: Weibo, crawler, users recommendation, information acquisition

1. INTRODUCTION

In 1967, Travers et al. proposed a theory of six degrees of separation through the chain letter experiment[1]. Dodds et al. tested this theory in a modern version of an email experiment in 2001 involving more than 60,000 volunteers from 166 different countries[2]. The results showed that the average complete email chain had only four people. Later, many scholars conducted a large number of relevant experiments on this theory and achieved certain research results[3-6]. The theory is that a person can get in touch with any stranger through no more than six intermediaries.

In recent years, the close combination of the "six degrees of space theory" and the Internet has begun to show commercial value[7]. People pay more and more attention to the research of social network, and many network softwares begin to support people to establish more mutual trust and close social connections, collectively known as "social software"[8]. The commercial potential of a trusted network generated by aggregation through six degrees of division is truly incalculable.

For example, Weibo[9] has established a micro social relationship network with new social media. With the popularity of Weibo, the communication among Weibo users is becoming easier and easier, and the Weibo system has already formed a micro-society[10]. Therefore, we can explore the practical significance of the six dimensions theory with the assistance of Weibo.

The purpose of this study is to establish a Weibo data acquisition and analysis system to help users find strangers who they want to know. We will be able to find a way to help us connect with our favourite celebrities.

2. SYSTEM FRAMEWORK DESIGN

In order to help Weibo users find their indirect friends, indirect network acquisition system based on Weibo user data is designed in this study. This exercise includes two subsystems: Weibo data acquisition system and Weibo data analysis system. The overall structure of the two systems is designed as follows.

2.1. Overall structure of Weibo data acquisition system

The end control condition of crawler is that when the attention information of Layer N (N optional Settings) friends of seed user is searched and saved, the program will stop running automatically. Figure 1 is the overall flow chart of Weibo data acquisition system.

*lizhizhuang3333@qq.com

International Conference on Computer Application and Information Security (ICCAIS 2021),
edited by Yingfa Lu, Changbo Cheng, Proc. of SPIE Vol. 12260, 122601S · © The Authors.

2.2. Overall structure of Weibo user relationship analysis system

The Weibo user relationship analysis system has three functional modules: database connection module, data display module and query module.

The database connection module is used to connect the Weibo user relational database. Data display module is used to obtain and display all user follower data in the database; The inquiry module consists of two sub-modules: contact inquiry module and middleman inquiry module. The interpersonal network query module can be used to query all indirect friends of a user, and the middleman query module can be used to query the intermediaries between a user and any stranger.

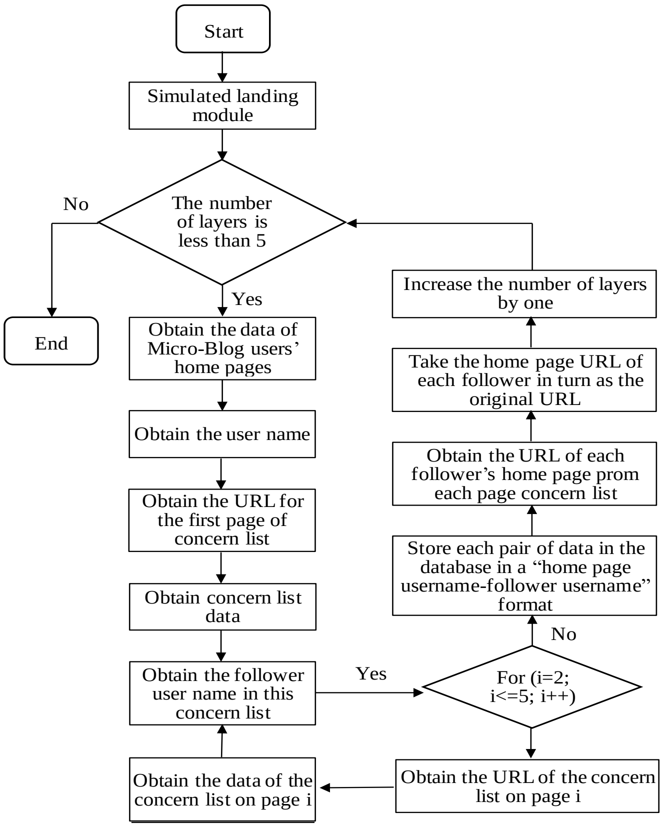
Start
Simulated landing module
The number of layers is less than 5
No
Yes
End
Obtain the data of Micro-Blog users' home pages
Obtain the user name
Obtain the URL for the first page of concern list
Obtain concern list data
Obtain the follower user name in this concern list
Obtain the data of the concern list on page i
Yes
For (i=2; i<=5; i++)
No
Obtain the URL of the concern list on page i
Store each pair of data in the database in a "home page username-follower username" format
Obtain the URL of each follower's home page prom each page concern list
Take the home page URL of each follower in turn as the original URL
Increase the number of layers by one

Figure 1. Flow chart of Weibo data acquisition system algorithm.

Figure 2 is the overall structure of Weibo user relationship analysis system.

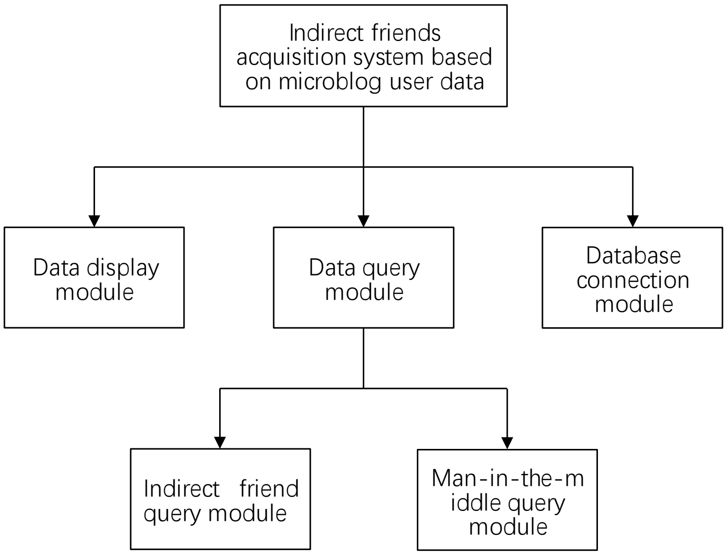
Indirect friends acquisition system based on microblog user data
Data display module
Data query module
Database connection module
Indirect friend query module
Man-in-the-middle query module

Figure 2. Overall structure of Weibo user relationship analysis system.

3. DETAILED DESIGN OF THE SYSTEM

3.1. Detailed design of each module of Weibo data acquisition system

3.1.1. Simulated Weibo recording module. Weibo needs under the condition of the login to access, so this topic first to study is to simulate Weibo record module implementation, this module requires the user to login the Weibo account through the program, so that after a successful login Weibo crawler can have access to other users' information through the HTTP address, such as following lists, follower lists and Weibo lists.

By analyzing the codes about logging in the web version of Weibo, it can be found that three steps of simulating Weibo login are needed to compile the module of simulating Weibo entry:

(1) The client sends a login request to the user server of Weibo.

(2) After receiving the login request, the server generates the corresponding secret key and returns it to the client.

(3) The client will combine the user name, password and the secret key sent by the server and submit the login information to the server. The server will return the login status and the user's personal information after successful verification.

After login, the client keeps a session with the server to access the data in the Weibo.

3.1.2. User name extraction module. The user name extraction module contains the following two functions: extracting the user name from the Weibo homepage and extracting the user name of the followers from the following list.

After the successful login of the Weibo crawler, it can access the relevant information of other users through HTTP address, such as other users' following list, follower list and Weibo list. After successful login, the crawler enters the homepage of the first Weibo user according to the initial URL. After entering the Weibo user homepage, the user name of the Weibo homepage and the user concerned should be obtained to prepare for storing Weibo relational data in the database in the future.

Since crawler uses HTTP request to obtain data, each request obtains a large number of complex HTML codes, but the user's Weibo data have the same format, and these data can be extracted from the chaotic HTML codes by regular expression. In Weibo, each user has its own unique ID corresponding to it. Therefore, ID can be used as the basis to identify the certain user. When access to the data of relevant users, only the user ID can be used to access the relevant data.

3.1.3. Concern list URL acquisition module. After obtaining and saving the correspondence between a Weibo user and all the other following users, the work of the crawler is not finished. The crawler should also take the homepages of all other users concerned by the user as a new starting point for a new round of data acquisition. The crawler needs to extract the keywords in the unified resource locator of each Weibo follower's home page from the following list and assemble it into a complete unified resource locator.

The function of this locator is the extraction of the keywords from the web page of a user's following list and the connection of the keywords and the specific characters into a complete web page of following list. This module can not only obtain the first page of the following list (the relationship list of the first page) but also the following list from the second page to the fifth page. Other users can't be allowed to scan the following list from the sixth page, except for the user.

Because the following number is different, the number of following lists is also different. If the page number of following list may not be more than five, in case of errors, the keywords should be extracted from the limited following list, rather than the list of mandatory five pages. This module determines how many pages of following list the current user has through a judgment function, so that a scheduling function can classify the users with more than five pages of following list and the users with less than five pages.

After obtaining all the contents of a Weibo use's home page, this module uses the function "re.findall()" in Python's "re" library to search for character strings starting with "http:\/\/weibo.com\/p\/" and ending with "\" through regular expression matching. And it uses string length calculation function and string interception function to extract the 16 numbers in the middle part, then uses the string concatenation function to connect the fixed portion of the UNIFORM Resource Locator for the list of concerns we want to acquire the complete uniform resource Locator.

3.1.4. Weibo homepage URL acquisition module. The function of the module is to extract the keywords of the unified resource locator of all "followers" users' Weibo homepage from all the contents of the user's following list page, and connect the keywords and specific characters into a unified resource locator of the complete Weibo homepage.

After obtaining all the contents of the user's following list page, this module uses the function "re.findall()" in Python's own "re" library to match two different regular expressions, and extracts some strings in the middle part of the results of the two searches. And the two different string splicing methods are used to splice them together with two different fixed strings, so as to assemble a complete unified resource locator of Weibo homepage.

The first searches a string beginning with "/u\/" and ending with "\", and extracts letters or numbers from the middle part using string length calculation function and string interception function, and then uses string concatenation function to concatenate "http://weibo.com/u/" in front of it. We can obtain a complete Weibo home page of the unified resource locator.

The second searches the string beginning with "href=\"\/" and ending with "\", extracts the letters or numbers in the middle using the string length calculation function and the string interception function, and concatenates "http://weibo.com/" in front of it uses the string concatenation function. We can obtain a complete Weibo home page of the unified resource locator.

3.1.5. Database connection and data storage module. Using the MySQLdb interface from python to connect to the database, data storage can be performed. After operating all database, the database connection should be closed using cursor.close() and conn.colse() functions.

Every time Weibo users grab the required user relationship information through the relational capture module, the data storage module is called automatically. The module uses the connection object to obtain a CURSOR object, and then uses the method provided by the cursor to work. These methods include two categories: (1) execute a command; (2) receive a return value; The execution commands callproc(self, procname, args) can be used to execute stored procedures.

3.1.6. Scheduling function module. Only when each module of crawler is called in accordance with certain order and rules can the crawler run correctly and play its due function. The scheduling function module is a function used to control the operation of each module. In the system, functions are designed according to the program flow chart of crawler, and the functions of each module are called according to the sequence in the program flow chart.

3.2. Detailed design of each module of Weibo user relationship analysis system

3.2.1. Database connection module. The database connection module connects to the database with MySQLbd interface from python, so that the following operations can be performed.

3.2.2 Data display module. The module can obtain all Weibo user data by executing SQL data query statements using database operation functions and display the required Weibo data in the form of a list through dataGridView1_CellContentClick control.

3.2.3 Querying modules. The query module includes two sub-modules: interpersonal network query module and middleman query module. The function of this module is completed by SQL query statement.

The interpersonal network query module contains the function of searching interval middleman less than 5 indirect friends. For example, if A is follower of B, who is C's follower, then A is C's indirect network.

The intermediate friend query module can query the intermediate friend path between two people. For example, if A follows B, who follows C, and C is the follower of D, then the intermediary path between A and D is A-B-C-D. The number of intermediaries is regarded as the length of the intermediary path. Then, for the specified two people, the module can give all intermediary paths less than 5 lengths according to the ascending order.

4. CONCLUSION

In order to help Weibo users find their indirect network, this study designed a Weibo data acquisition and analysis module system. The module system firstly obtains data and constructs database through Weibo data acquisition module system, and then realizes the search function of indirect network and relative middlemen of specified Weibo users through Weibo user relationship analysis module system. The module system can help users find out the strangers who they want to communicate and the people among them, thus giving us the opportunity to find a feasible way to meet our favorite celebrities or other public figures.

REFERENCES

[1] Travers, J. and Milgram, S., "Experimental study on the small world problem," Sociometry, 32(4), 425-428 (1969).

[2] Dodds, P. S., Muhamad, R. and Watts, D. J., "An experimental study of search in global social networks," Science, 301(5634), 827-829 (2003).

[3] Wang, S., "Whose is six degrees of space?," Internet World, (2), 1 (2008). (in Chinese)

[4] Jiang, K. and Sun, Q., "Social network service usage analysis of college students (Series II)," China Education Network, (1), 27-29 (2013). (in Chinese)

[5] Liu, H., Lu, H., Zhang, N., Zheng, X. L., "Theory research based on microblog of six degrees space," Application Research of Computers, 29(8), 4 (2012). (in Chinese)

[6] Lan, Y., [Research on Communication Social Relationship Network Based on Six Dimensional Space Theory], Sun Yat-sen University, Guangzhou, Master's Thesis, (2009). (in Chinese)

[7] Hu, X., [Research on Network Topology Generation Algorithm Based on Six-Degree Segmentation Theory], Shanghai University, Shanghai, Master's Thesis, (2014). (in Chinese)

[8] Ferguson, D. A. and Greer, C. F., "Using twitter for promotion and branding: An analysis of the content of local television twitter sites," Journal of Broadcasting, 32(13), 315-316 (2013).

[9] Pan, Y., "Network news communication in Web2.0 era," Journalism Lover, 21(4), 37-38 (2009). (in Chinese)

[10] China Internet Information Center, [The 33rd Statistical Report on the Development of Internet in China], China Internet Network Information Center, Beijing, (2015).