Behavior Analysis of College Students on Using Microblog in Big Data Environment

Peng Yang 1*, Wei Li 1

1 North China University of Technology, Beijing, 100144, China
*Corresponding author’s e-mail: yangpeng@ncut.edu.cn

Abstract. With the development of information technology in big data environment, the “closed” campus has gradually turned into an open campus, which brings a series of new opportunities and challenges for the student management of college. Social platforms such as Microblog can provide students with a broader space for learning and communication, students are no longer limited to being managed, and they are increasingly inclined to their own opinions. In this paper, a distributed Microblog spider platform is studied and implemented for the behavior of student users. First, the data support can be provided for behavior analysis by the crawling of user information and content of Microblog. Secondly, the combing, integrating, analyzing and mining of the crawled data can make the useful value behind the information fully utilized in the student management work.

1. Introduction
With the rapid development of mobile Internet, people’s attention to some social emergencies has been increasing, and the way of paying attention has gradually shifted from traditional newspapers and magazines, radio and television, etc. to Microblog and social networking sites. Netizens use the words on the Internet to convey information and also influence other netizens. Besides, the information spreads from mouth to mouth, so that a strong social force can be formed through this “information relay”. As a more active group of netizens, college students are increasingly publishing their opinions on social networks, sharing their lives and communicating with others. Therefore, social platforms have potential and far-reaching impact on college students who have not yet entered the society in the new era, which provides a new way for the education of college students [1]. As one of the most important social platforms in today’s society, Microblog publishes, comments and forwards millions of Microblogs every day [2]. Such a large and open social platform can generate amounts of information every day, which provides massive data support for public opinion monitoring. The analysis and mining of Microblog information can indirectly obtain the interests and concerns of students, which also provides reliable data support for the daily management of students.

Web Spider is a powerful program for automatically extracting webpages, which follows the users’ specified requirements without any user intervention [3]. But the crawling efficiency is very limited and cannot cope with exponential explosive data growth [4], which is solved by the emergence of a distributed web spider system. The distributed spider system combines multiple hosts to process the crawling task together. That is, the same crawling queue corresponds to multiple schedulers and multiple downloaders, which doubles the crawling efficiency. This paper designs and implements a master-slave distributed web spider platform for Microblog of college students, which supports the crawling of multi-class Microblog data. Besides, the nodes are easy to expand and can adapt to the large-scale growth of data. Therefore, it can be used for fast, stable, and scalable crawling of...
2. Scientific analysis of distributed Microblog spider platform

2.1. Feasibility analysis

2.1.1. Simulated login

In the unlogged-in state, the data provided by Microblog page is very limited, which is not conducive to our analysis of user behavior. By analyzing the relevant code of login part of webpage Microblog, it is found that the Microblog login is mainly divided into three steps:

- The client sends a login request to Microblog server;
- After receiving the login request, the server generates a corresponding key and returns it to the client;
- The client combines the username, password and login key returned by the server in second step and submits it as login information to the server. After the authentication of the server, the correct login status and the personal information of current user will be returned.

After successful login, the client only needs to maintain a session with the server to easily access some data resources in Microblog [5]. In the distributed Spider platform, the other nodes only need to obtain the cookie of the logged in node to implement the login operation after any distributed node performs the login operation.

2.1.2. Distributed task scheduling

Distributed tasks are divided into real-time tasks and timed tasks. Real-time tasks are added via send_task() to respond to new requests in a timely manner. The scheduled task extracts a task from the task queue for execution at regular intervals [6].

The distributed architecture is implemented based on Celery (as shown in figure 1), communication between distributed nodes is implemented by brokers, and each node performs distributed tasks as a worker. Besides, a host acts as a server to assign tasks to distributed nodes and to supervise the working state of each node. When the server detects that a worker hangs, it will try to resend the task information to other workers [7] to ensure the robustness of the system. Since Celery comes with task routing, we can perform different crawling tasks on different nodes according to the actual situation [8].

2.1.3. Webpage crawling

After the login is successful, the distributed node starts executing the assigned task. The webpage crawling module extracts the seed URL from the task and passes it to the URL queue to be crawled. Next, DNS resolution and download of the webpage will be performed. The URL of the downloaded webpage is saved in the crawled URL list to avoid the webpage being repeatedly downloaded [9], and the downloaded web content is saved in the downloaded web library. The design of webpage crawling module is shown in figure 2.
2.2. Efficiency and cost analysis
When the system is configured with 10 machines, the information crawling amount of the Microblog spider platform can reach one million in one day, which is satisfactory. Sina server limits the number of times that an IP can access the API per hour [10]. Therefore, the number of distributed nodes is more important than the performance of single-node machines. The platform only needs to configure a computer with better performance as the server for task allocation and scheduling.

3. Functions of distributed Microblog spider platform
The functions of the platform include: crawling of hot topics, crawling of user information, crawling of user Microblog and crawling of Microblog comments. The platform provides a GUI interface to configure the spider task (as shown in figure 3).
3.1. Function of hot topic crawling
This function searches for Microblog with keywords as input and crawls popular Microblogs containing the keywords. The crawled content includes: username, source of the publication and the content posted. Topic crawling with the school name as a keyword can effectively determine which school the student came from.

3.2. Function of user information crawling
The user information that the function crawls includes: nickname, gender, location, profile, registration time, authentication status, number of followers, number of fans and number of sent Microblogs. The “learning experience” in the user information plays a key role in determining whether the user is a college student.

3.3. Function of user Microblog crawling
This feature enables crawling of all Microblogs of a given user. Crawled content includes: publishing time, Microblog content, likes, comments and forwards. The picture is displayed as a URL, and the behavior analysis of college students on using Microblog depends on the crawled Microblog contents.

3.4. Function of Microblog comments crawling
This feature enables crawling of a specific comment on Microblog. The returned content contains: the reviewer ID, gender, nickname, comment time, comment text, location and likes. Crawling on student Microblog can effectively capture the alumni of the student, which is one of the effective ways to collect students from a specific school.

4. Experimental design and analysis

4.1. Experimental design
The main work involved in this experiment is: the acquisition of Microblog users of college students and the analysis of user behavior based on Microblog text content. The way to obtain the Microblog student users of a specific school is as follows:

- Crawl the comment information under the official Microblog of the school and save the user information of the comment;
- Filter the users of the comments and save the users between ages of 16-26; if the school is filled in the education information column of the user, treat them as students of the school, skip the next step;
- Perform the Microblog crawling function on the filtered user. If the name of the school's iconic location (dining hall, teaching building, apartment, etc.) does not appear in the user's Microblog, discard the user.
When the number of users is insufficient, the school name can be used as a keyword to crawl the hot topic. After the user information is obtained, the second and third steps can be performed, and the obtained student user can also be crawled by Microblog comments to mine the students of the same school in the review users.

The Microblog content published by user is an explicit expression of user’s interest. Since the Microblog refers to multimedia content such as pictures and videos, it is processed into plain text information before analysis. Behavior analysis is based on the publish time of Microblog and the areas covered by Microblog content. The keywords included in the Microblog are matched with the keywords in the thesaurus, and the matching degree between the Microblog and each thesaurus is counted. The five thesauruses used for analysis are: Star, Game, Food, Tourism and Study.

4.2. Experimental results and analysis

The experiment was conducted on the popular official Microblog of Tsinghua University, and 800 users who were studying in Tsinghua and aged between 16 and 26 were selected. By crawling the user’s Microblog, 67,964 Microblogs were obtained. The time for student to post the Microblog after the text analysis is shown in figure 4.a, and the field of Microblog content is shown in figure 4.b.

According to Figure 4.a, it can be seen that 06:00-08:00, 10:00-12:00 and 18:00-20:00 are the three most popular time slots for Microblog, which is in accordance with the phenomenon that students like to use Microblog during meals. At 02:00-04:00 and 04:00-06:00, there are still a certain number of Microblog posts, which indicates that some students are seriously staying up late, which should be paid attention to by our faculty. According to Figure 4.b, it can be seen that food has the greatest attraction to students, and tourism is also one of the main themes of Microblog content. This inspires us to replace the traditional rewards for outstanding students with food or tourism, which is a favorite way for young people in the new era.

5. Conclusion

Nowadays, college students have become a large group of Microblog users. The advantages of fast information update and communication speed make Microblog widely favored by college students, which has an important impact on correct outlook on life and values of college students. Through the application of distributed Microblog spider platform, the attention and use of Microblog of college students can be analyzed, which provides important decision-making basis for student management, thus helping teachers to carry out objective and rational thinking guidance to college students. Therefore, it plays a good role in promoting the personality training and value building of college students.

![Figure 4 Experimental results](image)

(a)Number of Microblogs sent by student users in each time period  
(b)Proportion of the field of Microblog content

References

[1] Chen, F., Bai, X., Zhan, S. (2014) A Study on Recursive Neural Network Based Sentiment
Classification of Sina Weibo. In: IEEE 13th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom). Beijing. pp. 681-685.

[2] Liu, B., Zhang, J., Qiang, L., et al. (2016) Research overview of microblog analysis. Journal of Hebei University of Science & Technology, 36:100-110.

[3] Nilambari G., N., Narendra M., S. (2016) A rule based intrusion detection system to identify vindictive web spider. In: 2016 International Conference on Computing, Analytics and Security Trends (CAST). Pune. 271-275.

[4] Dodig, B., Krstinic, D. (2005) Distributed web spider based on cluster and grid technologies. In: 27th International Conference on Information Technology Interfaces. Cavtat Dubrovnik. pp. 52-58.

[5] Li, B., Zhang, Y., Zhou, C., et al. (2016) A Microblog Data Acquisition Method Based on Follow-Group Mode. In: Ubiquitous Intelligence & Computing & IEEE Intl Conf on Autonomic & Trusted Computing & IEEE Intl Conf on Scalable Computing & Communications & Its Associated Workshops. Beijing. pp. 634-637.

[6] Cuevas, E., Cienfuegos, M., Rojas, R., et al. (2015) A Computational Intelligence Optimization Algorithm Based on the Behavior of the Social-Spider. Computational Intelligence Applications in Modeling and Control.

[7] Wang, Z., Yu J., Ying C., et al. (2015) Energy-efficient strategy of distributed file system based on data block clustering storage. Journal of Computer Applications, 35: 378-382.

[8] Gao, Q. Design of Distributed Crawler Framework Based on Redis. (2017) Agriculture Network Information.

[9] Xie, D., Xia,W. (2014) Design and Implementation of the Topic-Focused Crawler Based on Scrapy. Advanced Materials Research, 850-851:487-490.

[10] Deng, Z., Fu, M., Zhang, Q. (2017) A Research of Time Slot Optimization Task Scheduling Strategy Based on the Hadoop. Journal of Northwestern Polytechnical University, 35:32-37.