ATLANTIS PRESS

Advances in Computer Science Research, volume 94

International Conference on Big Data, Electronics and Communication Engineering (BDECE 2019)

Research of Improved Microblogs Crawling Strategy Based on Time Feature

Hongyan Zhao[1], Jifeng Tian[1], Xin Ye[1] and Peiyu Liu[2]

[1]School of information engineering, Shandong Yingcai University, Jinan, China

[2]School of information science and engineering, Shandong Normal University, Jinan, China

***Abstract*—For the problem of real-time about the existing Microblogs page crawling strategy on getting the latest news, this paper proposes an improved Microblogs crawler crawling strategy based on time behavior. For Microblogs page having fast update speed, this strategy adds the time feature tag to the fetching URL, and when you grab this URL again, Comparing the URL time feature and the content of the Microblogs page, and comparing the correlation analysis of the same URL content of different time tag, Thus, it could improve the real-time performance of Microblogs information. The experiment results show that the improved MicroBlog crawler crawling strategy has better real-time than the existing Microblogs page crawling strategy in fetching information, which could reflect the latest change of Public opinion trend.**

Keywords—time feature; microblogs; crawling strategy; real-time

With the rapid development of the Internet, especially the mobile Internet, social media, especially as social media for creating, sharing and exchanging Chinese content, has made rapid progress. Microblog is a microblog. It is a widely used type of social media. As a sharing and exchange platform, it focuses more on timeliness and discretion. Currently, there are mainly Sina Microblog, Tencent Microblog, and NetEase Microblog, Sohu Microblog, etc., of which the most rapid development is Sina Microblog. According to Sina Microblog's 2016 financial report, as of March 31, 2017, Sina Microblog's monthly active users reached 340 million, including many industries, agriculture, government, commerce, etc., as Sina Microblog users, and it has rapidly become people's access, and an important platform for publishing related information, and thus Microblog also gathered a large amount of data.

I. INTRODUCTION

As a new important asset, how to find the hidden value behind the data is particularly important. The mining of Microblog data is mainly applied in the following aspects: (1) Through Microblog communication, influence the image of the enterprise and guide the consumer's business behavior; (2) Manage customer relationship with consumers through Microblog, improve service level and enterprise income; (3) Through data mining methods Discover potential business models or the latest public opinion changes [1].

The most important aspect of discovering the latest changes in public opinion through data mining is how to obtain these data. As an important tool for automatically obtaining relevant information from web pages, the crawler determines the value of microblog data mining in terms of the quality and efficiency of data capture [2-4].

Due to the nature of Microblog's web pages, the frequency of different Microblog page updates varies greatly, especially when there are unexpected events, the time characteristics of microblog information are particularly important. The more real-time information is, the more valuable it is. When acquiring microblog webpage data, it is necessary to adopt a suitable webpage update crawling strategy to ensure the timeliness of information capture. In the current microblog web crawling strategy, the duplicate webpage URLs adopt a de-emphasizing strategy, which improves the efficiency of webpage crawling, but may not be able to obtain the latest news in time in the crawling process and grasp the latest public opinion dynamics [5] In view of this, this paper adopts an improved crawling strategy for microblog web crawlers. For the feature that microblog web pages are updated faster, a time characteristic mark is added to the fetched URL. When the URL is crawled again, the URL is compared. The time characteristics and the content of the microblog webpage are compared with the correlation analysis of the same URL content subject with different time stamps, thereby improving the real-time performance of crawling information. Experiments show that the improved Microblog webpage crawling strategy is better than the current Microblog webpage crawling strategy in obtaining real-time news, and can more effectively reflect the latest changes in the public opinion situation, thus winning more for guiding and controlling the development of Internet public opinion.

II. TRADITIONAL WEB CRAWLERS

The web crawler mainly refers to a program that crawls web resources by accessing the web, analyzes the captured resources to obtain links, and then roams web pages pointed to by other links.

The web crawlers are divided into three categories according to their implementation technologies and architectures: general web crawlers, topic web crawlers, and deep web crawlers. In practical applications, these crawler technologies often combine with each other. The crawling of generic web crawlers has a certain degree of globosity and blindness, and it is crawled from the perspective of the entire network. The crawling of the theme web crawler is based on

certain needs and topics. Its greatest feature is the ability to analyze content and determine the relevance of topics. The design of the in-depth web crawler has two parts more than the normal crawler form analysis and page status. By analyzing the web page structure and classifying it as an ordinary web page or deep web with more information, more web pages can be obtained, with certain intelligence and initiative [4-7].

Web crawling generally follows certain strategies based on the size of the topics and data captured. The following describes several strategies [7-8]:

A. The Strategy of Priority Crawling

A webpage contains many links. Therefore, after extracting links to web pages, how to continue crawling other webpages, there are two kinds of "order" problems that crawlers can choose to handle. Depth-first and breadth-first strategies. In the grab strategy, the breadth-first strategy is more preferred.

B. The Strategy of Eliminating Duplicate URLs

To achieve the goal that does not repeat crawling, crawler must remember the URL it has crawled. Only by remembering the past can it not be repeated. The huge scale of the links challenged the way and structure of recording. Whether it is possible to record a large number of accessed data with limited resources is a subject.

C. The Strategy of Webpage Priority Crawling

The priority crawling strategy of a webpage is to grab the high-impact webpages as much as possible while crawling, and to take care of high-impact webpages. The importance of a web page is determined by the two aspects of link popularity and link importance.

D. The Strategy of Webpage Revisit

The crawler keeps crawling the webpage, but the crawled webpage may have an update, so the crawler has to periodically refresh and revisit those webpages that have already been downloaded. By revisiting those updated webpages, the changes to the World Wide Web can keep pace with the times.

E. The Strategy of One Time DNS Parsing

To crawl a webpage, you first need to obtain the host IP address and port number based on the known URL. Then you can establish a socket connection and send and receive data. In principle, the parsed URL is parsed every time to obtain the IP address, but many web pages are under the same site. Therefore, the files on the remote host may use the same IP address, so only one parsing is needed; it solves the inefficiency of DNS repetitive parsing.

III. IMPROVED MICROBLOG WEBPAGE CRAWLING STRATEGY

On the mainstream microblog platforms (twitter, Sina Microblog, Tencent Microblog, etc.), regardless of the user's UID, or blog post, the commentary MID has a unique representation in digital form. And you can piece together the corresponding URL through the unique identifier of these digital types and log in to your personal homepage or blog post page. The traditional webpage crawling method can also be applied to microblog platform data crawling.

The work flow of Microblog web crawling is shown in Figure 1 [9].

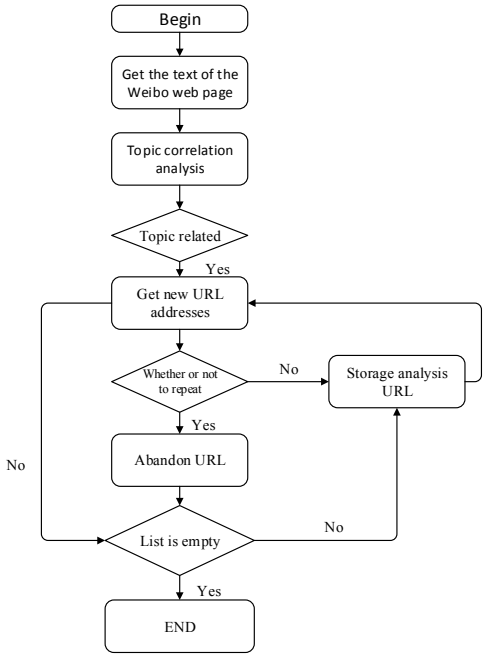

FIGURE I. MICROBLOG CRAWLER WORKFLOW

To crawl the Microblog webpage, the first to simulate the login process to log in, after the landing is successful, the microblog information crawling work, mainly including access to microblog page html code, analysis to obtain Microblog data, analysis to obtain Microblog user information and topics Relevance analysis, etc. Finally, the storage of relevant data includes the analysis of the subject related microblogs for publishing Microblog user information and the like.

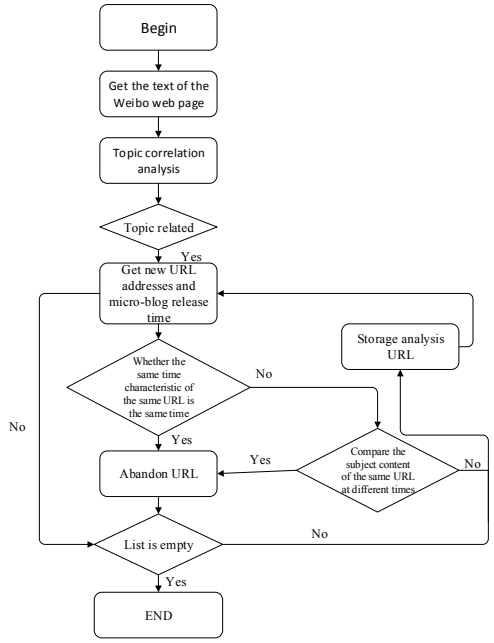

FIGURE II. IMPROVED MICROBLOG WEBPAGE CRAWLING STRATEGY

IV. EXPERIMENTS

A. *Experiment Environment*

1) Development Tools and Integrated Programming Environments: In the view of the demand of microblog web crawler, this paper uses C# as programming language and uses Visual Studio 2012 as the integrated environment for programming.

2) Relational Database System: Currently widely used database systems include Oracle, MS SQL Server, My SQL, MDB2, etc. Different application sizes and application scenarios will use different database systems. This paper uses Oracle as the relational database system.

B. *Microblog Website Crawler Strategy Analysis*

This paper uses the Sina microblog as the main object of crawl. The microblog grasps corresponding microblog data according to the given theme. When grasping the data, the system at the same time will grab microblog data and user information stored to the database.

In this paper, typical events on microblog are used as experimental samples, including Shandong ignominious murder (2017-03-27 10:03),Setting up new district in Xiong'an (2017-04-03 09:43), North China pollution incident (2017-4-18 15:30), Shandong Weihai school bus accident (2017-05-10 09:56), Bitcoin blackmail (2017-05-15 09:12). As the theme - based microblog web crawler method, we crawled of data for 3 hours on microblog. Comparing the improved strategy of crawler with the previous strategy:

1) Microblog fetching speed: The previous strategy based on microblog web crawler way crawled of data about Shandong ignominious murder (Event 1), Setting up new district in Xiong'an (Event 2), North China pollution incident (Event 3), Shandong Weihai school bus accident (Event 4), Bitcoin blackmail (Event 5). The amount of microblog data is 21150, 23310, 18540, 16650, 24930 respectively, while improved strategy based on the theme of microblog web crawler way gets 20430, 22110, 18030, 15630, 23070 respectively. In contrast, the pre-improvement strategy is more efficient at the speed of crawling on microblog.

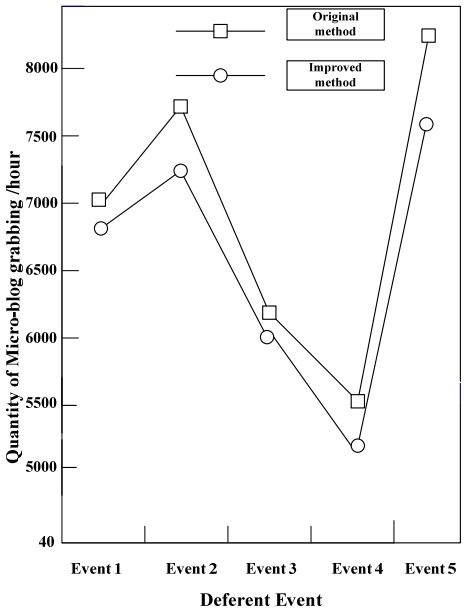

FIGURE III. FETCHING SPEED OF MICROBLOG

2) Real-time microblogging: Through the analysis of the microblog data captured before improvement, we study the development path and development situation of the event, including Shandong ignominious murder (Event 1), Setting up new district in Xiong'an (Event 2), North China pollution incident (Event 3), Shandong Weihai school bus accident (Event 4), Bitcoin blackmail (Event 5). We find that the improved strategy got the better real-time information. The time distance between the data development trend and the actual occurrence of the event is about 60min, 40min, 55min, 70min, and 30min respectively. In contrast, the improved strategy got the time distance is about 100min, 70min, 80min, 150min, and 45min respectively.

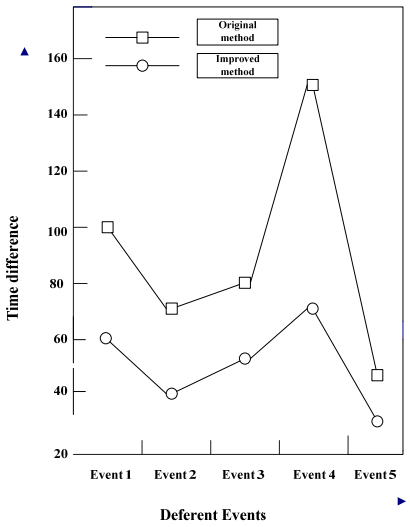

FIGURE IV. REAL-TIME PERFORMANCE OF MICRO-BLOG GRABBING

Therefore, relevant conclusions can be drawn from the above experiments and analysis:

1) In this paper, we study the microblog crawler strategy based on the improved crawler way while in microblog data fetching speed slightly behind the pre-improved strategy, but for the microblog event in the development of real time perform stronger, more easy to grasp of microblog event burst tendency perfectly, the change of the trend of public opinion more accurately, so as to guide and control the development of network public opinion has more room.

2) During the development of these five typical microblog events, the event of Bitcoin blackmail has a direct impact on the Internet. The result is the biggest impact on microblog users. Therefore, it is best to get the most data on the Internet, and the real-time performance of microblog is the best, and the actual data analysis is the smallest time difference between data development trend and event occurrence.

V. CONCLUSION

Based on the study of working process of the microblog network crawl, we put forward an improved microblog crawl strategy based on the time features, in view of the existing microblog page fetching strategy for the latest news on real-time problems. The strategy according to the characteristics of microblog page which updates fast, the fetched URL to add time characteristics of tag, when grabbed again to this URL, comparing the URL time characteristic and the microblog web content, comparing different time to mark the same URL content subject correlation analysis, thus improve the real time information of crawl microblog. The experimental results show that the improved microblog website crawler strategy can better reflect the latest changes of public opinion trend than that of traditional microblog. The next step in the microblog web crawl, considering the quantity of readers and more attention to the number of "Sina V", took the information in the climb, focuses on the influence of the "Sina V" on the related topics, thus more accurate judgment of development momentum.

ACKNOWLEDGMENT

This work was financially supported by the project of Shandong Social Science Project (16CFXJ18, 18BJYJ04).

REFERENCES

[1] HAN Jiawei, KAMBER M. Data Mining Concepts and Techniques[M].Ming Fan, Xiaofeng Meng, et al. Beijing: Mechanical Industry Press, 2001:120-336.

[2] Bin Liu, Jingyuan Zhang. Research Overview of Microblog Analysis[J]. Journal of Hebei University of Science and Technology, 2015, 36(1), 100-110.

[3] Hongjing Lin, Mengxing Huang. The Key Word Library Crawler Strategy based on Microblog Information[J]. Journal of Hainan University Natural Science Edition, 2016, 34(2):112-120.

[4] Jingjing Liu. Research and Implementation of Web Crawler for Microblog[M]. Shanghai: Fudan University, 2012.

[5] Mingjie Zhang. Design and Implementation of Bruising Data Acquisition System based on Web Crawler Technology[J].The Programming Case, 2015, 6: 72-75.

[6] Yishu Luo. Research on the Related Technology of Microblog Crawler[M]. Harbin: Harbin Institute of Technology, 2013.

[7] Zhao Hongyan. Research on Detection Algorithm of WEB Crawler[J]. I ntenerationl Journal of Security and Its Applicaitons,9(10),137-146.

[8] Jie Lian, Xin Zhou, Wei Cao. Sina Microblog Data Mining Scheme[J]. Journal of Tsinghua University Natural Science Edition, 2011, 51(10): 1300-1305.

[9] Yueguang Dan. Research and Implementation of Key Technology of Network Public Opinion based on Microblog[D].Chengdu: University of Electronic Science and technology of China, 2013.

[10] Xiaohu Zeng. Research on the Microblog Website Crawler based on Theme[D]. Wuhan: WuHan University, 2014.