ATLANTIS PRESS
*Advances in Computer Science Research, volume 63*
International Conference on Artificial Intelligence and Engineering Applications (AIEA 2016)

# An Information Extracting Scheme for Netdisk

Jinkui Dong, Hua Zhang

State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing, 100876, China.

**Abstract.** The netdisk not only supplies users with plenty of storage, but also provide the facility to share data. Therefore it has brought great convenience to users, but it also becomes the important channel for malicious and infringing applications to spread. Due to the extraction of netdisk information will get into trouble with the input of CAPTCHA, it's difficult to detect the applications. In this paper, we propose a method to extract downloading links of netdisk and another method based on the Drop Fall Algorithm to undertake the recognition of merged characters. We combine the numbers of characters and minimum values to confirm the initial points. The performance analysis shows that the method of extracting link is effective, and the recognition of characters is suitable for the netdisk environment.

**Keywords:** Netdisk; link extract; CAPTCHA; merged character; Drop Fall Algorithm.

## 1. Introduction

With the development of Internet, the way people get the information has changed dramatically. At present, people usually get useful information through search engines, yet the place that stores maximum information is cloud storage. Netdisk is one of the applications of cloud storage that is the most frequently used. In the field of information retrieval, web crawler is one of the fundamental components of search engine. However, because of the complexity of link extraction of netdisk and the netdisk's restriction of single IP address, the common crawler can't get the information you wanted.

Currently the research of web crawler is focused on fetching strategy [1], distribution technology [2], the algorithm of focused crawler [3] and the deep web crawler [4,5] . In the field of engineering, the research is focused on the design of frame [6, 7] and the combination of search engineer and web crawler, such as the Scrapy which is coded with Python and the Nutch which is coded with Java. However, it's inefficiency to fetch the netdisk information and even they can't get the netdisk information in some condition. The current crawler is not suitable for the information extraction in netdisk which is different in general HTML webpages. The real downloading links of files stored in netdisk are dynamic created when users click the link rather than directly saved in the webpages. Furthermore, after several times of crawling the program is required to input the CAPTCHA.

There are several difficulties for netdisk crawler to pass the anti-crawler mechanism, such as (1) the extraction of hidden downloading link and (2) the limit of downloading times in short time period. In this paper, we use program to imitate the operation of web browser and construct the requests to get the downloading links of files in netdisk to solve the former problem. Then we use the Drop Fall Algorithm that combines the characters number and minimum value to confirm the initial spot to ensure the recognition rate of the CAPTCHA of the netdisk and improve the success rate of downloading files stored in netdisk in short time period with single IP address. The performance analysis shows that the method of extracting link is effective, and the recognition of characters is suitable for the netdisk environment.

In next section, we present the analysis of extraction of downloading link. The recognition of CAPTCHA is illustrated in section 3. The section 4 shows the performance analysis and the experiment. The last section concludes this paper.

## 2. The extraction of downloading link

There two effective methods to solve the problem of the extraction of hidden downloading links in the current research:

1) Using program to call the functions offered by web browsers, and simulating to click the downloading button in the webpage to get the final downloading link, and downloading the file in program.

2) Using capture tool to analyze the network communication, and using program to simulate the communication and construct the related requests to get the real downloading link, and downloading the file in program.

The first method is inefficient, since the functions offered by the web browsers may result to download many css files and javascripts and occupy a lot of network bandwidth. The web browser will spend lots of CPU time to execute these files and scripts, which are not helpful to extract the final real downloading link. Therefore we select the second method to solve the extraction of downloading link which is effective.

In this section, we select a shared link 'http://**********/s/1o6Mqljc' as the example to undertake the analysis.

(1) Using Chrome or Firefox (or other web browsers with developer tools sets) to analyze the HTTP communication after clicking the downloading button. After filtering the HTTP communication that used to get javascript, css and picture, three useful HTTP communication message are reserved with an extra analysis.

1) The first is a POST request with a 200 response code. Different with general post requests, this one's required parameters contains two parts. One part is placed in the url, the other is placed in the form of the request. The request header contains the common field, including Host, User-Agent, and Accept etc. The url contains several parameters, including sign, timestamp, bdstoken, channel and web. The form submitted by this request has five key-values, including encrypt, product, uk, primaryid and fid_list. The response of the request is a json string.

2) The second is a GET request with a 302 response code. The url of this request derives from the dlink value of the json string of the first one's response.

3) The last is a GET request with a 200 response code. This request is the final real downloading link. It derives from the Location field of response header in the second request.

(2) Logid is the integrant parameter under the domain "***.*****.com" when requesting html and json.

Logid is a string composed by characters and numbers, and it ends up with '=', for example, "MTQ2OTMzMDk2NjcwAjAuMjQzODgzNzcwODg5MTg0LTY=" and "MTQ3MjcxOTYzMjM1MjAuNTEyNzQwMzczMDczMTU4Mg". The value varies every time when the parameter is used. Due to the confusion of javascript, it's difficult to find the source of this parameter. However, it can only be a random value or a confirmed value.

## 3. The recognition of CAPTCHA

The limit of the netdisk is that the user must input the CAPTCHA on the website when downloading many times in short time. It can be solved by two methods below:

1) Getting the upper limit of downloading times, then put up the distributed crawler to avoid the input of CAPTCHA

2) Using program to recognize the CAPTCHA and pass through the check of crawler

The best way to solve the problem is the combination of 1) and 2) in order to get maximum downloading times in practice. The first method is a matter of the resource, so this paper aims at the second method.

Using one netdisk as an example, by clicking the link 'http://**********/s/1o6Mqljc' in the web browser many times, a popup window will be displayed to user. The picture is shown in Fig. 1.


Fig. 1 the CAPTCHA of some netdisk

CAPTCHA (Completely Automated Public Turing Test to Tell Computers and Humans Apart) [8]. The CAPTCHA is a picture 40*120 pixels in size that is composed of 26 letters and 10 digits which are distorted and merged in this netdisk. The non-merged characters can be segmented by using vertical slicing algorithm [9], but the merged characters cannot be segmented by using vertical slicing algorithm, so the result of recognition is poor.

In this section, we use Drop Fall Algorithm [10] and the method of non-merged characters segmentation to segment the characters. Then, Tesseract-ocr [11] is used to recognize the characters. The process of segmentation of CAPTCHA characters is presented as follow. At first, CFS algorithm [12] is used to determine the character block. Then the number of block and the width of the blocks are used to determine the characters number of any block, and the vertical histogram projection is undertaken on the block with merged characters to get the minimum value of projection. At last, the characters number and the minimum value are combined to determine the initial point of Drop Fall Algorithm.

Among the process mentioned above, binarization algorithm [9] is used to pre-process the picture. The CFS algorithm [12], vertical projection algorithm (to get the minimum value) [8], Drop Fall Algorithm (for segmentation) [10] are used on the segmentation process. The binarization can set the background to 0 and the characters to 255, which can be easily distinguished by later process. The binarization algorithm can be divided into two categories: locality-based thresholding selection and global-based thresholding selection [13]. We select Ostu's Thresholding Algorithm which is one of the global-based thresholding methods [9].

The CAPTCHA contains four characters, but they are not all merged together in some pictures. Therefore, the method of segmentation on non-merged characters is applied on the picture. After the application, the original picture is divided into one to four regions. Because the total number of characters is certain, after the application of CFS algorithm [12] on the CAPTCHA, the characters number of every region need to be determined. There are four kinds of situations as below:

1) If there are 4 regions, every region contains 1 character.

2) If there are 3 regions, the widest region contains 2 characters and the others both contain 1 character.

3) If there are 2 regions, the width of the 2 regions need to be compared. If the width of the two regions is equal, the two regions both contain 2 characters respectively. If not, let us suppose the width of the two regions is width1, width2 and width1 > width2.If width1 > 2*width2, the region whose width is width1 contains 3 characters and the other contains 1 character. If not, the two regions both contain 2 characters respectively.

4) If there is 1 region, the region contains 4 characters.

If the characters number in the region is more than one, all the minimum value of the region need to be found as the reference point of Drop Fall Algorithm. The determination method of minimum values refer to this paper [10].

The original Drop Fall Algorithm [10] has many initial points, so there are multiple segmentation modes. However, the character numbers vary in 36 characters, which is much larger than 10 that the original application occasion has. The method is unsuitable for the environment. In this paper we propose a new method to determine the initial point of Drop Fall Algorithm. The method assumes that the region that need to be divided contains N (N=2, 3, 4) characters and the width of the region is W. Then two pointer starts to scans from the width that equals i*W/N(i=1, 2 …, N-1) to different sides at the same time until arriving on the closest minimum value point, and the point is regarded as one

initial point of Drop Fall Algorithm. Finally, N-1 points are found. After getting the initial points, the later process can be executed as the Drop Fall Algorithm does. We select Descending-Right algorithm [10].

## 4. Experiment

In this section, we collect 100 shared links of some netdisk, and succeeded to download 92 shared files with the methods mentioned above. The total size of the files is 2,104,513,018 bytes and it costs 3 hours and 6 minutes, and the mean downloading speed is 184KB/s. The speed can satisfy the demand of downloading the file about hundreds of megabytes.

Moreover, we downloaded 100 CAPTCHA pictures of some netdisk. It is consisted of 39 pictures that human eyes couldn't recognize and 61 pictures that could be used in this experiment. The result of segmentation and the recognition is shown in Table 1.

Table 1. The CAPTCHA pictures that human eyes can recognize

| the number of characters that are segmented accurately | the number of the CAPTCHA pictures | the total number of characters | the number of characters that are recognized accurately |
|---|---|---|---|
| 4 | 14 | 56 | 35 |
| 3 | 3 | 12 | 6 |
| 2 | 17 | 68 | 18 |
| 1 | 15 | 60 | 7 |
| 0 | 12 | 48 | 0 |
| Total | 61 | 244 | 66 |

The accuracy of segmentation of 4 characters 14/61=22.95%. The experiment segmented 4*14+3*3+2*17+1*15=114 characters accurately in total, and the total success rate of segmentation is 114/244=46.72%. The total characters number of recognition accurately with the Tesseract-ocr is 66, and the success rate is 66/114=57.89%. In conclusion, the success rate of segmenting 4 characters is comparatively lower, and it is only 22.95%.As to the recognition, the success rate is 57.89% which is well for a packaged program.

## 5. Conclusion

In this paper, we research the information extraction of netdisks and use the information extraction of some netdisk as an example to experiment. The results show that the success rate is 92%.For the input of CAPTCHA after lots of downloading (this experiment has encountered the CAPTCHA when downloading the 100 shared links in the last an hour), the preliminary recognition is undertaken. The accurate recognition is high in the offline situation and it can meet the primary demand.

## Acknowledgments

This work is supported by NSFC (Grant Nos. 61300181, 61502044), the Fundamental Research Funds for the Central Universities (Grant No. 2015RC23).

## References

[1] YU J, LIU Q. Survey on topic-focused crawlers [J]. Computer Engineering & Science, 2015, 2: 007.

[2] PU Q. The Design and Implementation of a High-Efficiency Distributed Web Crawler [C]. Dependable, Autonomic and Secure Computing, 14th Intl Conf on Pervasive Intelligence and Computing, 2nd Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress (DASC/PiCom/DataCom/CyberSciTech), 2016 IEEE 14th Intl C. IEEE, 2016: 100-104.

[3] Wang W, Chen X, Zou Y, et al. A focused crawler based on naive bayes classifier [C]. Intelligent Information Technology and Security Informatics (IITSI), 2010 Third International Symposium on. IEEE, 2010: 517-521.

[4] Bangar R, Kahate S, Scholar P G. New Approach for Web Crawler Using Data Mining to Discover Deep Web [J]. International Journal of Engineering Science, 2016, 6509.

[5] YU Cheng-long,YU Hong-bo. Research on Web Crawler Technology [J]. Journal of Dongguan University of Technology, 2011, 18(3): 25-29.

[6] Ahmadi-Abkenari F, Selamat A. An architecture for a focused trend parallel Web crawler with the application of clickstream analysis [J]. Information Sciences, 2012, 184(1): 266-281

[7] Bahrami M, Singhal M, Zhuang Z. A cloud-based web crawler architecture [C]. Intelligence in Next Generation Networks (ICIN), 2015 18th International Conference on. IEEE, 2015: 216-223

[8] XINGGUO LI, WEI GAO. Segmentation method for merged characters in CAPTCHA based on Drop Fall Algorithm [J]. Computer Engineering and Applications, 2014, 50(1): 163-166.

[9] Tingre S, Mukhopadhyay D. An approach for segmentation of characters in CAPTCHA [C]. Computing, Communication & Automation (ICCCA), 2015 International Conference on. IEEE, 2015: 1054-1059.

[10] Congedo G, Dimauro G, Impedovo S, et al. Segmentation of numeric strings [C]. Document Analysis and Recognition, 1995. Proceedings of the Third International Conference on. IEEE, 1995, 2: 1038-1041.

[11] Information on:https://github.com/tesseract-ocr.

[12] Yan J, El Ahmad A S. A Low-cost Attack on a Microsoft CAPTCHA [C]. Proceedings of the 15th ACM conference on Computer and communications security. ACM, 2008: 543-554.

[13] Yongying G, Li Z, Guowei W. An Algorithm for Threshold Based on Arithmetic Mean of Gray Value [J]. Journal of Image and Graphics, 1999, 6.