Journal of Engineering and Applied Sciences 13 (Special Issue 3): 3215-3220, 2018
ISSN: 1816-949X

Implementation of Real-Time Data Collection System Based on Improved Web Crawling Engine

[1]Ki-Bok Nam, [2]Koo-Rack Park, [3]Jin-Young Jung and [4]Han-Jin Cho
[1]Department of Computer Engineering,
[2]Department of Computer Science and Engineering, Kongju National Univercity,
32588 Chungnam, Cheonan Subuk, 1223-24 Cheonan-daero, Budeadong 275, South Korea
[3]Department of Bio Information, Daejeon Health Science College, 34504 Daejeon,
Dong-gu, Chungjeong-ro 21, South Korea
[4]Department of Smart Mobile, Far East University, 76-32 Daehak-gil, Gamgok-myeon,
Eumseonggun, 27601 Chungcheongbuk-do, South Korea

Abstract: Today, artificial intelligence has become an issue in the overall society. To learn artificial intelligence, it is required to collect big data. The conventional web crawling technique for data collection sometimes fails to read web pages when web sites change or servers shut down and has no any separate algorithm implemented in case of calling a page late. Therefore, the method has difficulty with data accuracy and effective data collection. In order to analyze the HTML DOM structure of a particular website to collect data and obtain proper data, this study implemented java based web crawling engine. Also, the implemented engine has an alarm function which is used to give an administrator the notification of the problem of a relevant website and uses multithread for fast simultaneous collection of multiple data. A relevant site may consider such an action to be a DDoS attack. Therefore, to solve the problem, the implemented engine does not access the same URL. When the system proposed in this study was applied, it shortened the time to obtain about 440,000 images at Google website from 58 h to <4 h. With the implemented system, it was possible to obtain data simultaneous by multiple keywords. Therefore, it is expected to collect big data more easily and accurately for artificial intelligence learning. In the future, the proposed web crawling engine will be developed further and therefore the system to process and classify collected data will be researched.

Key words: Web crawling, big data, multithread, DDoS, HTML DOM, AI

INTRODUCTION

Since, the industrial revolution, today's society has rapidly been developed. In other words, from the 1st industrial revolution in the late 18 century to the current 4th industrial revolution of ICT, remarkable development has been achieved. The 1st-3rd industrial revolutions had mostly focused on machine based physical areas for productivity improvement, whereas the 4th revolution concentrates on the connection of machines and ICT for productivity improvement and makes possible production and operation with accurate data (Ahn and Lee, 2016; Han and Kim, 2017).

Among the innovate technologies of the 4th industry, artificial intelligence is a reasoning system that uses big data for learning to draw a proper result in order to run on its own in the connection of IT and a machine. Of the techniques of big data collection, the typical techniques include the internet of things sensor based data collection and the web crawling based web data collection (Ki-Bok *et al*., 2017; Lee and Jang, 2016).

A conventional web crawling technique fails to read a relevant web page because of website change, server shutdown and environmental factors or call a web page late. To solve the problems, the web crawling engine implemented in this study uses a separate algorithm to send automatic monitoring results to an administrator so as to collect big data quickly and accurately.

Literature review

DDoS (Distribute Denial of Service attack) detection technique: Attack detection technique is divided into pattern analysis and probability analysis. In pattern analysis, the patterns of the packets used in UDP (User Datagram Protocol) flooding or TCP_SYN (Trasmission Control Protocol Synchronize Sequence Numbers)

Coreesponding Author: Koo-rack Park, Department of Computer Science and Engineering, Kongju National Univercity,
32588 Chungnam, Cheonan Subuk, 1223-24 Cheonan-daero, Budeadong 275, South Korea

flooding are saved in DB with the technologies of data mining like intrusion detection technology and then the packets received in the same pattern are detected (Lee *et al*., 1999). The pattern analysis is able to detect 100% the same pattern but is almost unable to detect any attacks modified, though slightly.

The detection technique using SNMP (Simple Network Management Protocol) DDoS uses MIB (Management Information Base) as a set of management objects to collect traffic and applies a threshold value to analyze attack traffic. This method consists of traffic collection step and analysis step. In traffic collection step, SNMP is activated between a management system and a target system and the MIB (Management Information Base) to manage is selected, so as to obtain information (Mun-Su and Chang-Seok, 2008; IL-Jun and Tae-Yong, 2014).

Web crawling: As a computer software technology, web crawling means the extraction of information from websites. It is a program of visiting internet web pages and collecting data. When visiting a web page, the program also visits other pages linked in the web page. Since, the program crawls around web along links in the same way as a spider, it is called 'Spider'. It is almost impossible to collect and categorize a massive amount of web documents manually. For this reason, in web document search such action is performed automatically.

For instance, let's assume that it is necessary to obtain the current music ranking data at Melon website. It may be manually possible to visit the site, check one by one, make an excel file and report to a boss. However, to use this data in a different program or for something other things, it is necessary to save the data in a DB. In different categories such as ranking, release date, title, album title and group name there are a variety of information.

A web page is basically created in a HTML form. With the use of 'view page source' or 'check developer', it is possible to look at how the page information is generated in a HTML form. These sources are usually managed by developers in a certain structured form. For this reason there are certain rules. Web crawling is the action of analyzing the rules and obtaining proper information (Kim *et al*., 2011).

HTML DOM: DOM (Document Object Model) can be considered to be the defined structure of all factors constituting a web page screen.

In other words, it defines the internal structures of the factors on screen. If DOM is used, it is possible to make asynchronously, processed data accessed to a screen dynamically. Therefore, it plays a very important role in AJAX. In addition, if DOM is used it is possible to make a dynamic change and give accessibility and operation to a relevant page (So-Jin *et al*., 2003).

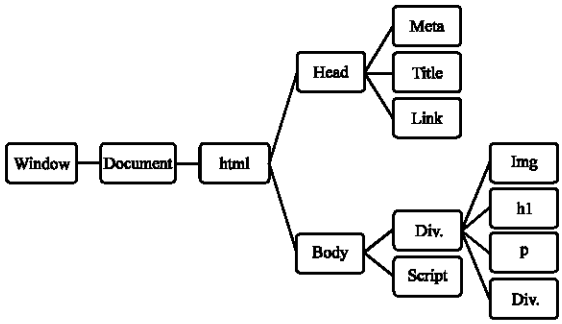

Fig. 1: DOM tree

DOM is able to integrate almost all spaces of HTML, including style and value in object for access and to operate all contents and their object values on screen. The model with such functions is called DOM tree. DOM NODE is one factor of a tree structure and is the most fundamental unit of DOM tree (Kim *et al*., 2013; Hyeon-Sook *et al*., 1992).

This method helps a program or a script access a document's contents, structure and style more effectively to update, replace and delete them. Figure 1 illustrates the structure of DOM tree.

MATERIALS AND METHODS

The proposed system: Figure 2 shows the architecture of the implemented system. The system proposed in this study is comprised of four managers. First, monitor manager receives the state of a called site and delivers an event to SMS manager when an issue occurs in order to give notification to an administrator.

Secondly, multi-thread manager is used to collect data simultaneously and manages threads to extract data from multiple different websites. Also, agent accesses the URL obtained from monitor manger, takes data of the website, delivers any change and state of the website to monitor manager and save an access pattern and URL in database.

Thirdly, WWW manager supervises web basic information including title, URL, access time and response time and delivers the HTML DOM and resource of the completely accessed site to agent.

Fourthly, DDoS manager regularly monitors Database to prevent the same pattern and same access URL, analyzes a pattern and instructs access and pattern change to WWW manager. Figure 3 illustrates the flow chart of the script that searches for an internet website in the implemented system. To collect the latest information of a particular website, URL is called and the URL

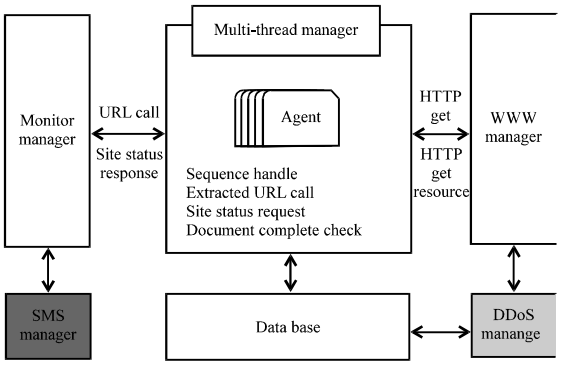

Fig. 2: Architecture of the proposed system

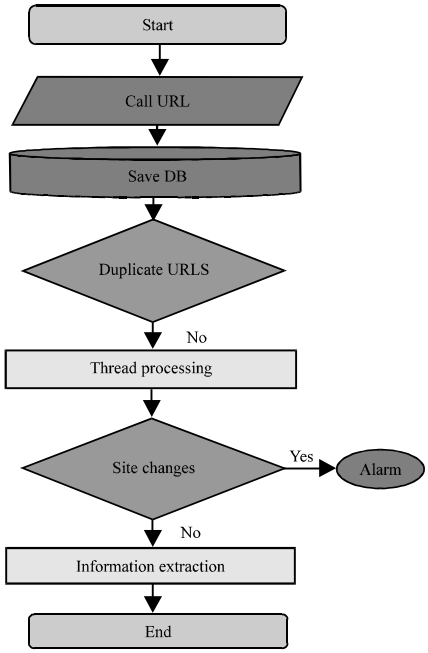

Fig. 3: Flow chart of the script searching for an internet website

information is saved in DB so as to prevent simultaneous access. The log of information collection date is recorded. The system makes a check to prevent duplicate execution as in DDoS and processes threads to collect information from multiple websites simultaneously.

It, also finds a site change or failure to access a site and send a text message as an alarm to an administrator, so as to collect the latest information with no problem. If there is no site change, the information extracted from the final site is collected.

In case of the obtainment of particular information from a particular website, rather than random site information extraction in general search, a website can be changed frequently and consequently it is impossible to extract proper data. To solve the problem, the algorithm presented in Algrothim 1 gives an administrator notification of the relevant contents and constantly obtains the latest information of a website.

Algorithm 1; Algorithm for extracting the latest information:

```
Smslist = AlramSms (“010-1111-2222, 010-7777-8888”)
CallUrl (“http://www.naver.com”)
Skip = CheckingTagN (“a”,”class = nhk”)
if (!skip)
{
    admin_sending (smslist)
]
Function AlramSms (phone num)
if (phone_num<>”“)
{
    phone_nums = phone_num.split (“,”)
]
return phone_nums
End function
Function CallUrl (baseUrl)
    driver_sebu = driver_sebu_m.getPage (baseUrl)
    wait (driver_sebu, “”)
End Function
Function checkTag (tag,chkhtml)
    List<HtmlElEment> dis_atag = ((HtmlPage) driver_sebu).getElement
ByTagName (tag)
    int skip_cnt = 0;
for (HtmlElement a_tag_item:dis_atag)  {
    Dchk = chkhtml.split (“=”)
    if (a_tag_item.getAttribute (Dchk[0]).inde)   ×Of (Dchk[1])  > -1) {
        rtn = true
        break
    ]
    else
    {
        rtn = false
    ]
]
    return rtn
End Function
```

The proposed algorithm monitors the state of a website and gives an alarm to an administrator in order to collect the latest information and to prevent any trouble of information collection.

First, AlramSms function makes a setting to send a text message to an administrator if a website has a problem.

Secondly, CallUrl is the function to obtain website information. In the function, wait function is additionally implemented in order to execute the next command after page information is fully loaded.

Thirdly, CheckTagN function is used to check if a website is accessible or if a website is changed. Therefore, there is no problem with the collection of the latest information.

Fourthly, CheckTagN function is used to send a text message to multiple administrators simultaneously if some problems like website access problem occur.

RESULTS AND DISCUSSION

Figure 4 presents the Google website to search for images. To test the implemented system, it is necessary to apply many data. Therefore, these researchers used the image search function at Google website in order for AI image learning and crawled images with particular keywords to extract data. In the Google image search, it is required to scroll down the screen to take images in the next page.

By checking the lastly obtained image is equal, it is necessary to find if there are more new images actually. If not, it is required to bring the URL of each image in the next step. Unfortunately, because of Google security, a URL can be hidden or is encoded with Base 64 code.

The Encoded screen is then Fig. 5 which makes it difficult to acquire images. Therefore, to solve the problem and bring the images in actual websites, this implemented system generates an automatic click for each image to obtain a relevant URL as shown in Fig. 6 Also, the duplicated images were excluded.

Algrothim 2 illustrates the screen of obtaining data with the use of the web crawling engine implemented on the basis of the contents in Fig. 4 and 6.

Algrothim 2; The captured image of the web crawling engine based data obtainment:

==image_keyword=watermelon**Start
==image_url=https://cookieandkate.com/image/2012/07/how-to-make-watermelon-juice-1.jpg
==image_url=https://peopledotcom.file.wordpress,com/2017/07/watermelon_ripe_hero_getty_getty.jpg?w=2000
==image_url=https://cookieandkate.com/image/2012/07/how-to-make-watermelon-juice-1-jpg
==image_url=htts://s3-us-west-2.amazonaws.com/beachbody-blog/uploads/2015/06/Mintry-Watermelon-Shakeology-2.jpg
==image_url=https://peopledotcom.file.wordpress,com/2017/07/watermelon_ripe_hero_getty_getty.jpg?w=2000
==image_url=https://cdn6.bigcommerce.com/s-hql52gp4vk.products/441/image/892/watermelon_10ml_88294.1405333614.1280.1280_35958.1495734 716.500.750.png?c=2
==image_url=https://s3-us-west-2.amazonaws.com/beahbody-blog/uploads/2015/06/Minty-Watermelon-Shakeology-2.jpg
==image_url=https://peopledotcom.file.wordpress,com/2017/07/watermelon_ripe_hero_getty_getty.jpg?w=2000

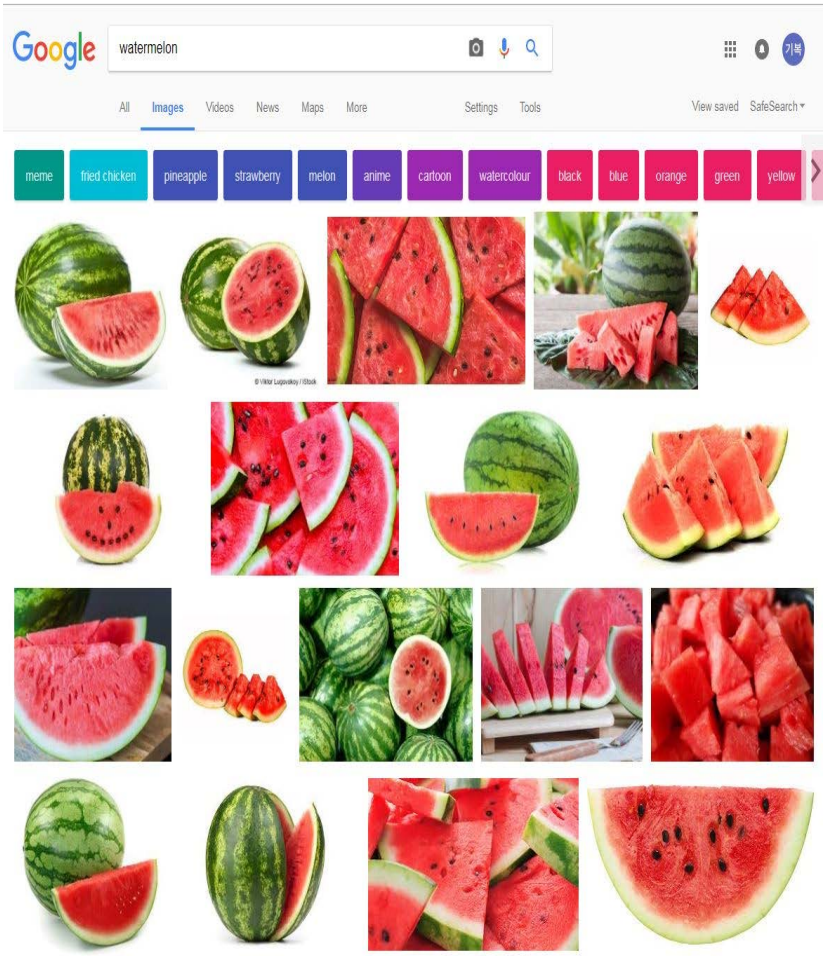

Fig. 4: Google image search

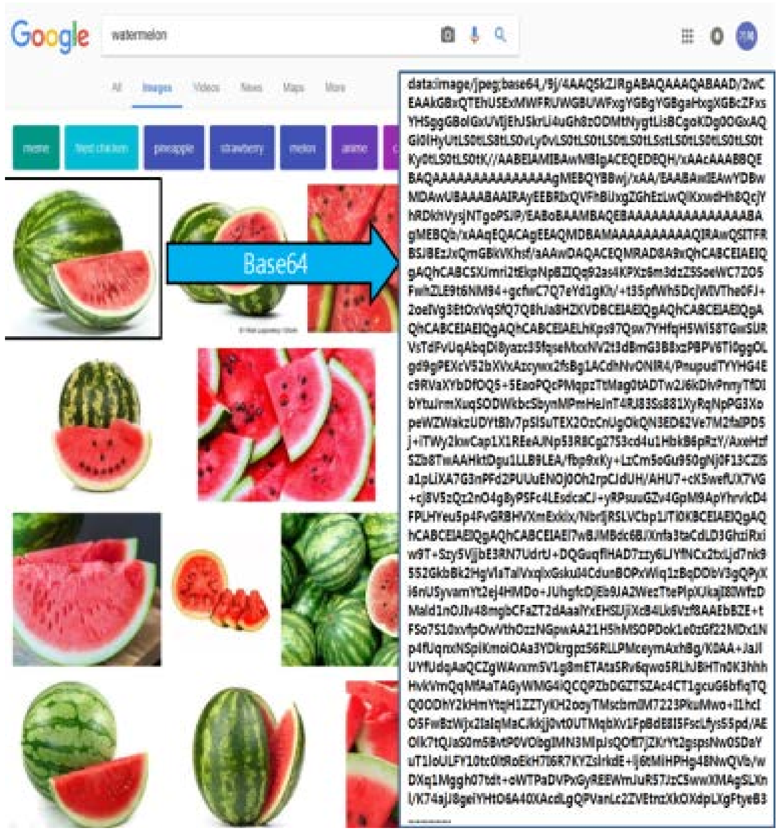

Fig. 5: Base 64 encodeing

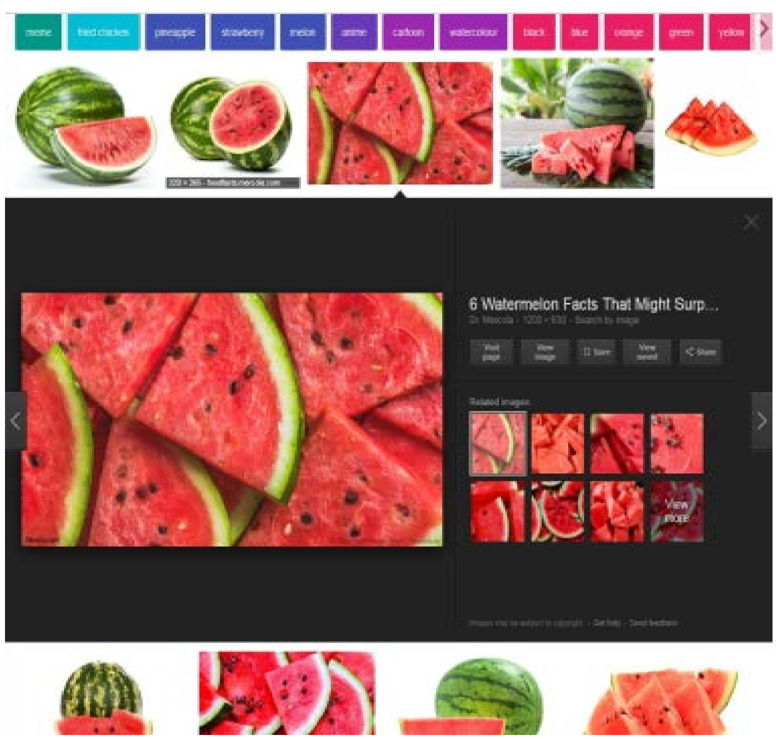

Fig. 6: Image advanced searc

On average, it takes about 140 msec per page or 0.14 sec per page to read one page and extract image data. Therefore, to bring about 440,000 data, it takes about 58 h. When fourteen threads were used simultaneously in the proposed system, it took about 4 h to extract data.

CONCLUSION

A conventional web crawling technique fails to read a relevant web page because of website change, server shutdown and environmental factors or call a web page late. To solve the problems, a new web crawling engine was implemented in this study. The implemented engine was used to obtain Google image data. As a result, it shortened the time to obtain about 440,000 images at Google website from 58 h to <4 h. Also, the implemented system was able to obtain data on the basis of multiple keywords. Therefore, it is expected to collect big data for AI learning more easily and accurately.

RECOMMENDATIONS

In the future, these researchers will collect image data on the basis of multiple keywords with the use of the implemented web crawling engine and will continue to study the implementation of image processing and automatic classification engine in order for AI learning of the collected data.

REFERENCES

Ahn, S.H. and M.H. Lee, 2016. Fourth industrial revolution impact: How it changes jobs. Korean Acad. Soc. Bus. Administration, 1: 2344-2363.

Han, H.S. and H. Kim, 2017. 4th Industrial revolution and knowledge services. Keit Ltd., Oxfordshire, England, UK.

Hyeon-Sook, L ., J. Jae-Hong and B. Doo-Kwon, 1992. A study on the generation of Dynamic Overview Map (DOM) in hypertext. Korea Inf. Sci. Soc., 1: 1039-1042.

IL-Jun, C. and S. Tae-Yong, 2014. Mix of traffic for detection of DdoS attack detection techniques. Korean Inst. Inf. Technol., 1: 232-235.

Ki-Bok, N., K.R. Park, C. Young-Suk, K. Joon-Yong and Y. Myung-Seob, 2017. A study on web crawling improvement model for real-time data collection. Proceedings of the 7th International Conference on Convergence Technology Vol. 7, July 5-8, 2017, ICCT, Hokkaido, Japan, pp: 1108-1109.

Kim, K., J. Park and B. Kim, 2013. Effective method to change multimedia scene configuration information using DOM update. J. Broadcast Eng., 18: 43-58.

Kim, K.Y., W.G. Lee, M.H. Lee, H.M. Yoon and S.H. Shin, 2011. Development of web crawler for archiving web resources. J. Korea Contents Assoc., 11: 9-16.

Lee, C. and J. Jang, 2016. Development of social data collection system using web crawling. Korean Inf. Sci. Soc., 1: 1787-1789.

Lee, W., S.J. Stolfo and K.W. Mok, 1999. A data mining framework for building intrusion detection models. Proceedings of the IEEE Symposium on Security and Privacy, May 9-12, 1999, Oakland, California, USA., pp: 120-132.

Mun-Su, J. and O. Chang-Seok, 2008. Prevent traffic analysis attacks by a web application. J. Korea Soc. Comput. Inf., 13: 139-146.

So-Jin, N., K. Do-Hoon, K. Wan-Jung and K. Yong-Hyuk, 2003. Dom-based content extraction for improving performance of web service. Korea Inf. Sci. Soc., 1: 218-222.