Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 157 (2019) 444–449

Procedia Computer Science

www.elsevier.com/locate/procedia

4th International Conference on Computer Science and Computational Intelligence 2019 (ICCSCI), 12–13 September 2019

Social Media Web Scraping using Social Media Developers API and Regex

Lusiana Citra Dewi[a,*], Meiliana[a], Alvin Chandra[a]

[a]*Computer Science Department, School of Computer Science, Bina Nusantara University, Jakarta, Indonesia 11480*

Abstract

Nowadays many information can be easily accessed through the internet. Some of social media web applications, such as Facebook, Twitter and Instagram, even provide user with an easy information sharing features. However, the information is presented in the form of a timeline or a feed, which is sometimes not relevant to the user or quite hard to accessed by user because of the redundancy of the information. This situation can be resolved with a Web Scraping method proposed in this paper, that able to search information, combine and present it in a better way according to user preferences. A system is developed to implement the proposed method by using an API that Facebook Developers and Twitter Developers provided. In addition, regular expression (or Regex) which is a language construction that can be used for matching text by using some patterns. Based on experiment conducted in this research, overload information could be suppressed into structure data that store in a database, less redundancy information is presented, and information relevancy could be adjusted to user preferences.

*Keywords:*Web Scraping; Regex; Facebook Developers API; Twitter Developers API Social Media;

* Corresponding author. Tel.: +62-812-9402-7121.
E-mail address: lcdewi@binus.edu

10.1016/j.procs.2019.08.237

1. Introduction

Technology developments are increasing rapidly. Many information is circulating on the internet. Nowadays, all information could be search easily. Search engine site like Google and Bing facilitate the information searching process. However, information provided by search engine site or social media application is sometimes overload. Users often find difficulties when they need information that is relevant to their needs. For example, to look for promotional information that is relevant to the needs of the user or events such as seminars that are in accordance with the field of the user.

Some of the social media that many of us already know, such as Facebook, Instagram and Twitter are one of best information spreading and searching facilities. Because it is not only providing the information itself, but also provide the information by users' social network. The information relevancy produced by social media application is at medium-high rate. It's a plus point for information spreading using social media. But there are also minus points, the information that are spread using social media are quite hard to search or reach the target audience at real time. It also causing difficulties in information extraction because there are too many redundancies of information when people start to post the same information.

Based on that background there are some problem formulation such as: (1) information overload, (2) information redundancy in social media, and (3) information relevancy to user.

What we need now is a system that can search the relevant information, filter this kind of information, and present it to user. Web scraping is one of the most popular technique to extract web data or content[1]. Facebook developers and Twitter developers provide an API that can allow our website (or application) access the information from the social media website. We have to make an AppID and AppSecret in order to use this API in our system or application. And then we make the regex pattern that user desire, and the match it with the raw information that we acquire with the API. The information that we get after matching the information and regex, are then stored to database for further usage, which is can be shown to the user. Based on this database we can sort the information based on its time, popularity, or relevancy.

2. Literature Review

Pereira et al.[2] state that Web Scraping is a process to extract data from internet with any method or technique. Web scraping also help web automation in many ways, including weather data, web changing detection, and price comparing online website.

Web Scraping can change unstructured data into a structured data that can be stored and validated into a database. The points of web scraping are collecting data, storing data, and validating data. The next important step is data analysis, thus validated data can be interpreted into a better information.

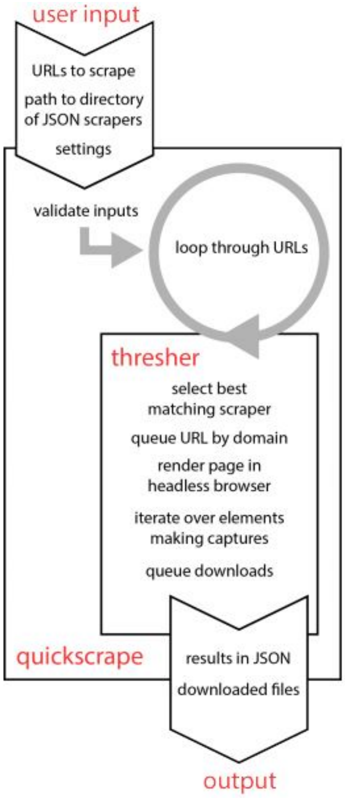

Fig. 1. Schematic overview of quickscrape of ContentMine Application

Richard Smith et al. [3] draw the schematic overview as shown on figure 1 above. Richard et al. [3] state that the need of data collection for academic purposes are very important. ContentMine is scraperJSON scrapers for major publishers. ContentMine also contain thresher, quickscrape, and another supporting library.

ScraperJSON uses JSON which this format is the most commonly used and in natural format of a JavaScript. It became a standard for scraping process. Thresher itself is a web scraping library that uses ScraperJSON and enabling it for headless browsing. Quickscrape is a web scraping command line using the thresher and also a supporting library to extract data and metadata.

Another research of Arbelaitz[4] show the usage of web mining technique to extract information from website and adapt the information for specific user's need. Moreover, Sidik[5] present the work to explain website structure redesign by using web mining technique (web usage mining). Web usage mining technique provides discovery and analyzing of usage patterns in order to concern the needs of web applications. This technique uses server log data to generate the user navigation pattern that will recommend website structure redesigning format. Flesca[6] also mines user preferences, page content and usage to personalize website navigation.

2.1. Facebook and Instagram Developers API

For using the Facebook and Instagram API we have to make a facebook account and then go to official Facebook Developers website (https://developers.facebook.com/) to register our application and then acquire the Application ID (AppID) and Application Secret (AppSecret). This ID are use further in the development of the web scraping application so we can send several requests to Facebook for data. AppSecret will be used to decode the encrypted messages from Facebook, so that sensitive information remains protected.

2.2. Twitter Developers API

Twitter Developers API works quite similar as Facebook Developers API. We must make a Twitter account and then go to official Twitter Developers website (https://developer.twitter.com/en.html) to register our application and then acquire the Application ID and Access Tokens.

3. Research Method

This research is started by analyzing the information flow in social media, how users search information through social media, and how the information is presented to user. Information extraction method is using web scraping that implemented by Facebook Developers API and Twitter Developers API. Information extracted will be matched with user preferences by using regular expression (or Regex) which is a language construction that can be used for matching text by using some patterns. Algorithm development phase will be including the snippet code design that can be used for web scraping.

4. Analysis and Results

4.1. Web Scrapping Algorithm

Web scraping process started with the choice of source, because it affects which API we'll be used, and then what authentication that will be needed for each social media website. Below is the snippet code example for authentication process for web scraping from Facebook:

```
//appid & appsecret can be found from https://developers.facebook.com/apps
string appid = "APPID";
string appsecret = "APPSECRET";

//get OAuthURL
string oauthUrl =
string.Format("https://graph.facebook.com/oauth/access_token?type=client_cred&client_id={0}&client_secret=
{1}", appid, appsecret);

//get AccessToken
string accessToken = client.DownloadString(oauthUrl).Split('=')[1];
```

Next process is to choose the web page that will be the target of scraping and determine number of data that will be scraped in the range of 0 until 100. Where 100 is the maximum number.

```
string content =
"fields=posts.limit(100){message,full_picture,created_time,updated_time,link,picture,permalink_url}";

try
{
        string pageInfo =
        client.DownloadString(string.Format("https://graph.facebook.com/{0}?access_token={1} ", target,
        accessToken));
}
catch
{
        Console.WriteLine("No User Found");
}
```

The data that API collect are on JSON formatted data. After that the data will be parsed using available method such as:

```
string pagePosts =
client.DownloadString(string.Format("https://graph.facebook.com/{0}?{1}&access_token={2} ", target, 
content, accessToken));
JObject Posts = JObject.Parse(pagePosts);
```

This kind of data that we will used to match with the regex format that we desire. With regex format we can detect any kind of information that we will need.

```
//REGEX SESSION
string revent = "promotion|event|seminar|concert";
for (int i = 0; i < dpKey.Length; i++)
{
        Match mobjects = Regex.Match(dpKey[i],revent,RegexOptions.IgnoreCase);
        if(mobjects.Success)
        {
                objects.Add("true");
        }
        else
        {
                objects.Add("false");
        }
}
```

After that the data can be inserted to database for further use.

4.2. Information Trend Algorithm

For acquiring trending information, we will need some counter variable, where it will count number of the redundant data.

```
privatestatic IEnumerable<KeyValuePair<string, int>> FindDuplicates(string[] array)
{
        Dictionary<string, int> stringSet = new Dictionary<string, int>();
        foreach (var item in array)
        {
                int count;
                if (stringSet.TryGetValue(item, out count))
                {
                        stringSet[item] = count + 1;
                }
                else
                {
                        stringSet[item] = 1;
                }
        }
        return stringSet.Where(p => p.Value >= 1);
}
```

And then we will update the field counter in the database with it, and sort descending it.

5. Conclusion

Based on the testing conducted using white-box testing methodology, the main function of the algorithm can be achieved, such as: (1) the information overload can be surpressed into structured data stored in database, (2) there are minimum numbers of information redundancy which that we convert into trend analysis data, and (3) relevancy of the information can be changed by changing regex format.

This research can be further develop into an application. Web application that can be accessed from anywhere or any devices will be better. And for further studies the web scraping algorithm can be develop with an automated or scheduled scraping, so that the business or application will running automatically.

References

1. Glez-Peña D, Lourenço A, López-Fernández H, Reboiro-Jato M, Fdez-Riverola F. Web scraping technologies in an API world. Briefings in bioinformatics. 2013 Apr 30;15(5):788-97.
2. Pereira RC, Vanitha T. Web Scraping of Social Networks. International Journal of Innovative Research in Computer and Communication Engineering. 2015 Oct;3(7):237-40..
3. Smith-Unna R, Murray-Rust P. The ContentMine scraping stack: literature-scale content mining with community-maintained collections of declarative scrapers. D-Lib Magazine. 2014 Nov;20(11/12).
4. Arbelaitz O, Gurrutxaga I, Lojo A, Muguerza J, Pérez JM, Perona I. Web usage and content mining to extract knowledge for modelling the users of the Bidasoa Turismo website and to adapt it. Expert Systems with applications. 2013 Dec 15;40(18):7478-91.
5. Khan S, Singh Y, Sharma K. Role of Web Usage Mining Technique for Website Structure Redesign. International Journal of Scientific Research in Computer Science, Engineering and Information Technology3. 2018;1.
6. Flesca S, Greco S, Tagarelli A, Zumpano E. Mining user preferences, page content and usage to personalize website navigation. World Wide Web. 2005 Sep 1;8(3):317-45.