Tracking Online Fraud Using Regular Expression

Fiftin Noviyanto, Dewi Soyusiawaty, Nur Rochmah
Dyah Puji Astuti
Department of Informatics Engineering
Universitas Ahmad Dahlan
Yogyakarta, Indonesia
fiftin.noviyanto@tif.uad.ac.id, dewi.soyusiawaty@tif.uad.ac.id,
rochmahdyah@tif.uad.ac.id

Rinaldi Munir, Masayu Leylia Khodra
School of Electrical Engineering and Informatics
Institut Teknologi Bandung
Bandung, Indonesia
rinaldi-m@stei.itb.ac.id, masayu@stei.itb.ac.id

Abstract—**The online fraud mode is still rife, either through Short Message Service (SMS) or website. The problems that exist are: many people cannot distinguish fraudulent and official websites, tracking sites by the police is less effective because it is only done one by one based on reported websites, whereas one fraud perpetrator can have multiple websites. Search websites on search engines require enormous effort, as not all of those websites are related to the fraud being tracked, shown on the first page. This research employs the Google Site Search Scrapping data in the form of Application Programmable Interface which allows offline data management, so that the search becomes more flexible and unlimited per page. The purpose of this tracking system is to track and find fraudulent websites based on certain keywords. The tracking procedure is adopted based on the Special Criminal Investigation Unit of Yogyakarta Regional Police Department. The keywords used for the input data is telephone number or website address. The telephone number is then filtered and break down into area code, space character, and number separator using regular expression (regex) method. The testing method of this study is done by using Software Usability Test method, which produced score 68.8. A score minimum for a system that can accepted are 68. So, based on the results of testing, known that the system is acceptable. The implementation of a tracking system it allows the police found some site false of one report, so to minimize the number of dupes.**

Keywords—fraud; tracking; regular expression; Google API

I. Introduction

According to the Head of Criminal Investigation Unit in South Jakarta Police Department, Adjunct Senior Commissioner of Police (AKBP) Audie Latuheru, they received 1-2 reports per day about fraud crimes during the year 2009-2010. While year 2011 until 2012, the number was raised to 2-3 reports per day. Reports are getting more increasing to 3-4 reports every day during this year [1]. A phone number used in a fraud case is also possible to be used on several websites at once.

The difficulty of fraud sites tracking led people to report fraudulent sites. This is because the process of checking by the police service is based on data reported by people affected by fraud. Tracking process is done by visiting the website one by one, as it considered less effective and spends longer time from the searching process until fraud is positively detected. In addition, limitations of technological skills of police officers also become a problem [2, 3].

It needs a bigger effort to search a website that is not found on the first page results since the ranking also depends on the Search Engine Optimization (SEO). These rankings are determined by the search engines algorithm [4]. And phone number can be written in several formats. So, it needs to do a repeated search for each telephone number.

It is known that Google provides a Uniform Resource Location (URL) along with its web content that is available online in the form of Application Programmable Interface (API). All things on the internet are indexed by search engines that it needs an enormous size of database. This also causes another problem when we want to find the desired website. In addition, phone numbers used in online fraud contacts can be written variously, for example: numbers with country code, numbers with 3-digit separation or by insertion of other characters. Those cases make search process become more serious because it requires to repeat different keywords for every search.

This research is conducted using the dataservice API with XML format which is provided publicly by the search engines to be used in the process of tracking. Regular Expression (RegEx) is a flexible and concise notation for searching and replacing text patterns, where the main function is to find and replace text patterns [5]. RegEx can identify a pattern of data that will be taken from the results of information retrieval from search engines or web scrapping automatically [6]. As a consequence, RegEx method may able to help police officers in retrieving fraudulent sites information from the internet by keyword expansion.

The implementation of web scrapping has been studied by Josi, et al. [7]. Their research resulted in a search engine application by applying web scraping technique to extract information of scientific journal articles from several academic portals from Indonesia and abroad, then store them into a database automatically. Unfortunately, the numbers used as the data input were not converted into various formats. As a result, the implementation of telephone number searching for scrap web information tracking cannot be done. This research develops a keyword splitter so that the search can be done once with various forms of phone number.

Regular expression can be used to solve various string-related problems such as done by Jia Liu and Huseng Liao [8] and [9] which conclude that polynomial time algorithm in XML literacy checking is more effective than the automated practical approach. Another study was conducted by Artur

978-1-5090-5864-8/17/$31.00 ©2017 IEEE

Backurs and Piotir Indyk [10] which concluded that the complexity of regular expression matching can be characterized by depth, while matching and membership testing can be resolved quickly.

This research attempts to apply expression online regular on fraud tracking system based on the number cell phones and address website. This research produces software web-based with 2 the level of users, the community as dupes, and investigators cops do tracking on the website associated with the password sought. Tracing made to the site of a web that was available at the base of google data through google site search fire. So, the application of a tracking system it reduces dupes online.

II. Materials and Methods

A. Materials

1) Cyber crime

There are two types of cybercrime [11]. The first type is a crime that targets computers, computer networks or electronic devices. There are four categories of this type, which is described as follows:

1. Unauthorized access offences, for example: hacking.
2. Malicious codes offences, for example: the spread of viruses and worms.
3. Interruption of services offences, for example: DOS (Denial of Service) attack that keeps the server busy.
4. Theft or misuse of services, such as theft, or using someone else's account.

The second type of crime is a crime that makes computers, computer networks and electronic devices a tool to commit crimes. This type is divided into 3 parts, which mentioned as follows:

1. Content violation offences, for example: pornography involving children, revealing secret military information, and IP violations.
2. Unauthorized alteration of data, or software for personal or organizational gain, for example online fraud. This section is the main focus of our research.
3. Improper use of telecommunications, for example: cyberstalking, spamming, use of services for the purpose of committing crimes.

2) Regular Expression

The input filtering process in this study is utilizing a Regular Expression method. Regular Expression (REGEX) is a language construct for text matching based on a particular pattern, especially for complex cases.

In addition, REGEX is also very powerful specifically in word parsing process. Regular expression allows us to search, substitute, or separate strings in complex cases [5]. Here is the Regex PHP function using PCRE and PCRE function syntax in PHP is displayed in Source Code 1.

```
1. int preg_match ( string $pattern , string $subject
   [, array &$matches [, int $flags = 0 [, int $offset
                     = 0 ]]] );
2. int  preg_match_all(  string  $pattern  ,  string
   $subject  [,  array  &$matches  [,  int  $flags  =
        PREG_PATTERN_ORDER [, int $offset = 0 ]]] );
3. mixed  preg_replace(  mixed  $pattern  ,  mixed
    $replacement , mixed $subject [, int $limit = -1 [,
                  int &$count ]] );
4. array preg_split ( string $pattern , string $subject
   [, int $limit = -1 [, int $flags = 0 ]] );
5. string preg_quote( string $str [, string $delimiter
                     = NULL ] );
6. array preg_grep string $pattern , array $input [,
   int $flags = 0 ] );
```

Source Code 1. The command to break the input format of the mobile phone number with the addition of characters

B. Methods

The steps taken to solve the problem of this online fraud tracking system include:

1) Requirement Analysis

To determine the system requirement, we conducted a study of literatures and observed the process of reports submission from the fraud victims. Users this system consisting of: 1) People who becomes the victims of fraud and 2) Police Department who tracks those fraud cases based on report, in the form of phone numbers and/or website address used by the perpetrator of fraud.

Analysis of main user needs and system requirements can be described as follows:

a) People

- The public can register on the system
- People can report online fraud based on telephone number or website address
- View the status of fraudulent reports

b) Police Department

- Police can see and track sites based on keywords, websites, mobile numbers from google search data.
- The police may change the status of progress reports

2) System Design

As general, the system architecture is shown in Fig. 1. This system will track fraud based on data input accomplished by the police officers. The process of tracking then be executed through the internet along with the website data which is indexed by Google as a form of dataservice.

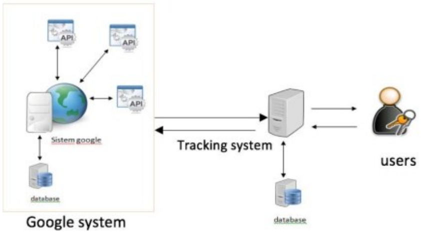


Fig. 1. Request data from the google system via the API.

Description of data request design from google system through API can be seen in Fig. 2. This paper completes the design inside line strip box. And Fraud filtering process will be completed in the future work.

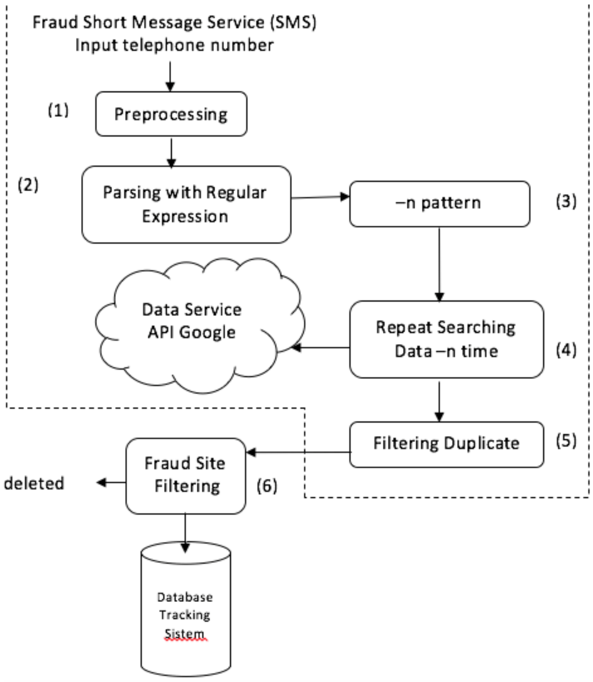


Fig. 2. Diagram of the process of tracking a phone number in Google's scraping data.

Process orders in Fig. 2 are described as follows:

(1) The process will start from inputting the fraud short message service (SMS) or phone number. The sequence for detecting phone numbers in sms strings, suchas:

a) Detection number in SMS string using command:

```
$get_str=preg_match('/([0-9]+)/',
$value);
```

b) Remove string on the number using command:

```
preg_replace('/[^0-9\']/','',
$value);
```

c) And then check the first number. Check the first number. If starting from +62 or 081 or 088 and another phone number format.

d) The number will be calculated whether 10, 11 or 12 digits. Where appropriate then it will be returned as telephone number.

SMS format including string and number phone. And the phone number format entered will be vary, including the addition of a country code, area code or a 3-digit separator character, for example: +62xxxxxxxxxx, 08xxxxxxxxxxxx, 08xx-xxx-xxx (x is assumed as numbers).

(2) The preprocessing phase include: filtering phone number from those various additional characters (spaces, characters, or area code). The detection command is displayed in Source Code 1.

```
function hp_num($nohp) {
$role_hp = array(' ','(',')','.','-');
foreach ($role_hp as $value) {
$nohp = str_replace($value,"",$nohp);
} $nohp =
str_replace("+62","0",$nohp);
return $nohp; }
```

Source Code 1. The command to break the input format of the mobile phone number with the addition of characters.

(3) The command in source code 1 is the regex code used to create a mobile phone number format based on the number of strings or numbers. Source code 2 is the code used to create a mobile phone number format to track the phone number of fraudsters, which will be inserted into array. As examples, the number formats that will be appeared are such as 0812-2345-6789, +628123456789, +62-812-3456-789, or + 62812-3456-789. This process will only search once, without repeating the search for the other format of numbers.

(4) Each type of number will be parsed using Regular Expression with the command in Source Code 2. Source Code 2 is the code used to create a mobile phone number format to track the phone number of fraudsters, which will be inserted into array. Tracking process will be performed on the dataservice provided by Google, while the retrieval process will be done with web scrapping. The data read will also be filtered with regular expression to detect the telephone number or just a regular number.

(5) The data will be filtered from duplicate searching result.

(6) Search results indicated by online fraud will be displayed. Once found the URL data of the website in accordance with the search, then the next stage is sorting to be stored into the database tracking system or not related to the case of fraud.

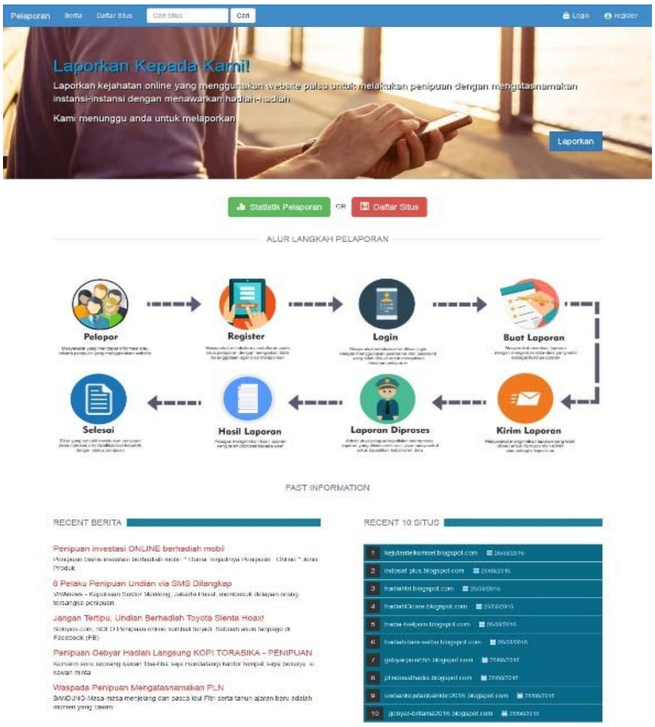
```
function hp_digit_split($hp){
if (strlen($hp) == 10){
 preg_match("/(\d{4})(\d{3})(\d{3})/",$hp,$o_num);
  $o_hp = "{$o_num[1]}-{$o_num[2]}-{$o_num[3]}";
        } else if (strlen($hp) == 11){
 preg_match("/(\d{4})(\d{4})(\d{3})/",$hp,$o_num);
  $o_hp = "{$o_num[1]}-{$o_num[2]}-{$o_num[3]}";
        } else if (strlen($hp) == 12){
 preg_match("/(\d{4})(\d{4})(\d{4})/",$hp,$o_num);
     $o_hp = "{$o_num[1]}-{$o_num[2]}-{$o_num[3]}"
} else { $o_hp = $hp;  }
return $o_hp; }
```

Source Code 2. Display the format of mobile phone number.

3) Implementation

System implementation stage is the process of altering the design and database forms into the programming language. We adapt HTML, CSS, JavaScript, JQuery, and Ajax JQuery used to build interface, while PHP language for processing data and MySQL for its database management system.

Fig. 3 depicts the reporting system layout view. The home page will display the flowchart for reporting information, recent posts, updated fraud sites and tips on handling fraud.



Laporkan Kepada Kami!

Laporkan kejahatan online yang menggunakan website palsu untuk melakukan penipuan dengan mengatasnamakan instansi-instansi dengan menawarkan hadiah-hadiah

Kami menunggu anda untuk melaporkan

ALUR LANGKAH PELAPORAN

FAST INFORMATION

RECENT BERITA

Penipuan investasi ONLINE berhadiah mobil

[illegible] Produk

8 Pelaku Penipuan Undian via SMS Ditangkap

VIVAnews - Kepolisian Sektor Menteng, Jakarta Pusat, membekuk delapan orang tersangka penipuan

Jangan Tertipu, Undian Berhadiah Toyota Sienta Hoax!

[illegible] Facebook (FB)

Penipuan Gebyar Hadiah Langsung KOPI TORABIKA - PENIPUAN

[illegible] kawan minta

Waspada Penipuan Mengatasnamakan PLN

BANDUNG-Masa-masa menjelang dan pasca Idul Fitri serta tahun ajaran baru adalah momen yang rawan

RECENT 10 SITUS

1 kejutantelkomsel.blogspot.com [illegible]
2 indosat plus.blogspot.com [illegible]
3 hadiahtri.blogspot.com [illegible]
4 [illegible]
5 [illegible]
6 [illegible]
7 [illegible]
8 [illegible]
9 [illegible]
10 [illegible]

Fig. 3. Display reporting system.

III. Results

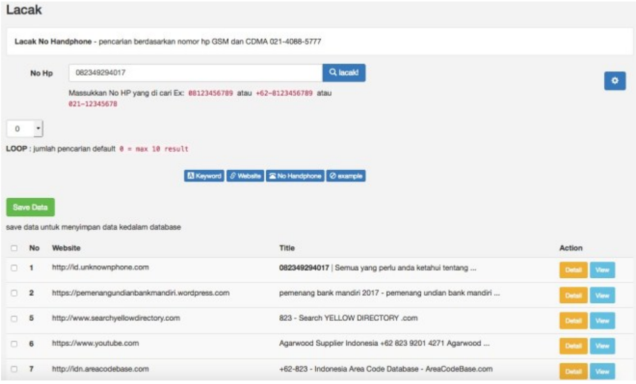
Tracking process based on phone number will be displayed on Fig. 4.

Testing phase was fulfilled by conducting Blackbox test and Software Usability Test (SUS). The Chief of the Special Crime Investigation Unit of the Yogyakarta Regional Police Department played role as a respondent to do the Blackbox testing.

The purpose of testing phase is to find out the conformity between the tracking process and the requirements so that the quality of the system can be guaranteed.

Based on the results of business function and process testing, it is concluded that the system is able to be used to assist the tracking process of online fraud cases.

For the Alpha test, there is an odd sequence Quote (positive sentence), which score is calculated on a minus 1 (xi-1) position scale.



Lacak No Handphone - pencarian berdasarkan nomor hp GSM dan CDMA 021-4088-5777

No Hp: 082349294017

Massukkan No HP yang di cari Ex: 08123456789 atau +62-8123456789 atau 021-12345678

0

LOOP : jumlah pencarian default 0 = max 10 result

save data untuk menyimpan data kedalam database

No	Website	Title	Action
1	http://id.unknownphone.com	**082349294017** | Semua yang perlu anda ketahui tentang ...	Detail View
2	https://pemenangundianbankmandiri.wordpress.com	pemenang bank mandiri 2017 - pemenang undian bank mandiri ...	Detail View
5	http://www.searchyellowdirectory.com	823 - Search YELLOW DIRECTORY .com	Detail View
6	https://www.youtube.com	Agarwood Supplier Indonesia +62 823 9201 4271 Agarwood ...	Detail View
7	http://idn.areacodebase.com	+62-823 - Indonesia Area Code Database - AreaCodeBase.com	Detail View

Fig. 4. Display tracked web page using phone number for keyword.

Questions with even order (negative sentence) works with rules as follows: scores are calculated at 5 minus / minus the position scale (5-xi). The overall SUS score is earned by contributing item scores to 2.5, so the overall SUS score is in the range 0-100 with the addition of or increment every 2.5 points.

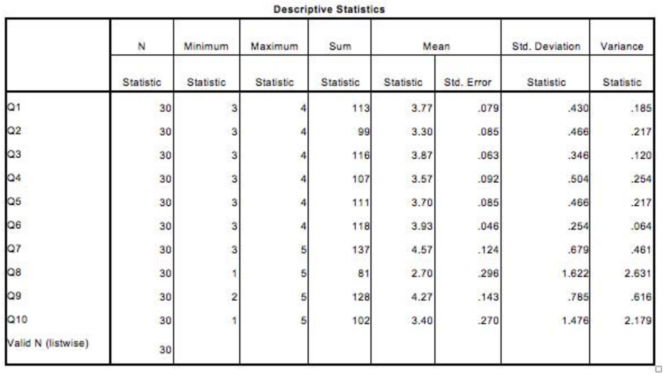
$$Total\ Score = \frac{\sum_{i=1}^{N} r_i}{N} \qquad (1)$$

Where r_i is the i^{th} respondent, N is the number of respondents.

From Table I we are able to determine the result of Alpha Test after utilizing System Usability Scale (SUS) method, while its results can be seen in Table I, and description analysis on Fig. 5.

TABLE I. The Result of Alpha Test after utilizing System Usability Scale (SUS)

Responden	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Average	SUS
1	5	3	5	5	5	3	5	5	3	5	4.4	55
2	5	5	5	3	5	3	5	3	4	4	4.2	65
3	5	3	5	3	3	5	5	5	4	5	4.3	52.5
4	5	3	5	3	5	5	5	4	4	5	4.4	60
5	5	3	5	5	5	5	5	4	5	5	4.7	57.5
6	3	5	3	3	5	5	3	4	5	4	4	45
7	5	3	5	3	5	5	4	5	5	4	4.4	60
8	5	3	5	3	3	5	3	5	5	3	4	55
9	5	3	3	3	5	5	4	5	2	3	3.8	50
10	5	3	5	5	5	5	3	5	5	5	4.6	50
11	5	5	5	3	3	5	5	5	4	5	4.5	47.5
12	5	3	5	3	5	5	4	3	3	5	4.1	57.5
13	5	3	5	3	5	5	4	3	3	5	4.1	57.5
14	3	3	5	3	3	5	5	3	5	5	4	55
15	4	3	3	5	3	5	5	3	5	5	4.1	47.5
16	5	1	5	1	5	1	5	1	5	1	3	100
17	5	2	5	3	5	2	5	1	4	3	3.5	82.5
18	5	1	5	1	5	2	5	1	4	1	3	95
19	5	2	5	3	5	2	5	1	4	2	3.4	85
20	5	2	4	2	4	2	5	1	4	2	3.1	82.5
21	4	2	4	2	4	1	5	1	4	2	2.9	82.5
22	3	2	4	3	4	1	5	1	4	2	2.9	77.5
23	5	2	4	3	4	2	4	1	5	2	3.2	80
24	5	3	4	2	5	1	4	2	5	3	3.4	80
25	5	1	5	1	5	1	5	1	5	1	3	100
26	2	1	5	3	4	3	5	2	4	3	3.2	70
27	4	2	5	2	3	2	4	2	4	5	3.3	67.5
28	4	2	5	2	3	2	5	2	4	3	3.2	75
29	3	2	4	1	4	2	5	1	5	1	2.8	85
30	4	1	5	3	5	1	5	1	5	3	3.3	87.5
Average	4.5	2.6	4.6	2.8	4.3	3.2	4.6	2.7	4.3	3.4	3.7	68.8
90%CI	0.26	0.33	0.20	0.34	0.25	0.51	0.20	0.49	0.24	0.44	0.18	4.99
Median	5	3	5	3	5	3	5	2.5	4	3	3.65	66.25



Descriptive Statistics

	N	Minimum	Maximum	Sum	Mean		Std. Deviation	Variance
	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Statistic
Q1	30	3	4	113	3.77	.079	.430	.185
Q2	30	3	4	99	3.30	.085	.466	.217
Q3	30	3	4	116	3.87	.063	.346	.120
Q4	30	3	4	107	3.57	.092	.504	.254
Q5	30	3	4	111	3.70	.085	.466	.217
Q6	30	3	4	118	3.93	.046	.254	.064
Q7	30	3	5	137	4.57	.124	.679	.461
Q8	30	1	5	81	2.70	.296	1.622	2.631
Q9	30	2	5	128	4.27	.143	.785	.616
Q10	30	1	5	102	3.40	.270	1.476	2.179
Valid N (listwise)	30							

Fig. 5. Description Analysis for SUS System.

IV. Conclusion

A tracking system cheating online use regular expression use the service of data from google fire has been served in this paper. There are two level users: community as reporter and police as user system to trace the cheating. Tracking through the phone number or address the site reported by the public. Next, do tracking system on the website indexed by Google search engine, through google search fire site. The implementation of this system is expected to facilitate the police to overcome fraud cases, so that minimize the number of fraud victims.

Further research to be done is to filter the sites that have been found based on the keywords entered. The filtering is based on the website indicated by fraud. The criteria of a website indicated website fraud is based on the rules in the police.

Respondents test over each 30 people and 10 questions. Questions given concerned with application and comfort when using system. Next, the results of the questionnaire analyzed by using the method testing SUS. Based on the average, SUS score of respondents reached 68.8 caused the system is stated in the acceptable categories. Minimum score accepted is 68.

Fig. 5. shows that all the questions to 30 respondents valid. Value minimum answer is 1, which means strongly disagree. While value maximum is 5, which means totally agree.

Application of regular expression in system l. Testing software by using the method alpha test against 30 respondents with 10 questions. The results of the analysis SUS 68.8 categorized as C, apply but can be developed. Further research will be increased based on the results of the low.

Acknowledgment

The research was funded by the General Directorate of Higher Education, the Ministry of Technology Research and Higher Education of Indonesia through Competitive Research Grant scheme 2017.

References

[1] S. Decilia, "Polisi Tangani 600 Kejahatan Online Per Tahun," *Tempo*, Apr-2013.

[2] D.W. Ismoyo, "Kendala penyidik dalam mengungkap tindak pidana penipuan online melalui media elektronik internet (studi di polres malang kota)," Universitas Brawijaya, 2014.

[3] I.M.A. Windara and A. K. Sukranatha, "Kendala dalam penanggulangan cybercrime sebagai suatu tindak pidana khusus," *Kertha Negara*, vol. 1, no. 4, 2013.

[4] H.S. Khraim, "The impact of search engine optimization dimensions on companies using online advertisement in jordan," *Am. J. Bus. Manag.*, vol. 4, no. 2, pp. 76–84, 2015.

[5] E. Spishak, W. Dietl, and M.D. Ernst, "A type system for regular expressions," in *FTfJP 2012: 14th Workshop on Formal Techniques for Java-like Programs*, 2012.

[6] R. Lawson, *Web Scraping with Python*. Birmingham, Mumbai: Packt Publishing, 2015.

[7] A. Josi and L. Abdillah, "Penerapan teknik web scraping pada mesin pencari artikel ilmiah," *Res. Gate*, no. September, 2014.

[8] J. Liu and H. Liao, "Intersection checking for regular expressions based on inference system," *Int. J. Database Theory Appl.*, vol. 8, no. 4, pp. 241–250, 2015.

[9] K. McGowan and B. Lagle, "A better way to search text : Perl regular expressions in SAS," USA, 2011.

[10] A. Backurs and P. Indyk, "Which Regular Expression Patterns are Hard to Match ?," in *2016 IEEE 57th Annual Symposium on Foundations of Computer Science Which*, 2016.

[11] I. Baggili, "Digital Forensics and Cyber Crime," in *International ICST Conference ICDF2C 2010*, Abu Dhabi, United Arab Emirates: Springer, 2011.