LONTAR KOMPUTER VOL. 10, NO. 2 AUGUST 2019 **p-ISSN 2088-1541**
DOI : 10.24843/LKJITI.2019.v10.i02.p02 **e-ISSN 2541-5832**
Accredited B by RISTEKDIKTI Decree No. 51/E/KPT/2017

Web Scraping and Winnowing Algorithms for Plagiarism Detection of Final Project Titles

Neng Ika Kurniati[a1], Alam Rahmatulloh[a2], Ridwan Nur Qomar[a3]

[a]Program Studi Informatika, Fakultas Teknik, Universitas Siliwangi
Siliwangi Street Number 24, Tasikmalaya City 46115, West Java, Indonesia
[1]nengikakurniati@unsil.ac.id, [2]alam@unsil.ac.id, [3]ridwan.nurqomar14@student.unsil.ac.id

Abstract

Plagiarism in research can occur due to accident or intentional. Plagiarism is an act that violates copyright and includes actions that harm others. In submitting the title of the research, for example, for the final assignment research, not a few students who repeatedly submitted titles were rejected and considered doing plagiarism because the title proposed had already existed before. Then we need a system that can detect the similarity between the titles to be submitted and the existing titles so that it is expected to reduce the occurrence of plagiarism. This study uses a winnowing algorithm to find the percentage similarity between titles. The Google Scholar will be used to obtain data on research titles that have been previously available as comparison titles. Web scraping with CURL (Client URLs) and simple HTML DOM parser is used to retrieve title data from Google Scholar. The results of the study with the application of a Winnowing algorithm to find the percentage similarity to data from Google Scholar were able to present a percentage of similarities in percent with the category of mild, moderate or severe plagiarism, while also helping early detection as prevention of plagiarism.

Keywords: *Final Project, Google Scholar, Plagiarism, Web Scraping, Winnowing Algorithm*

1. Introduction

Determination of whether or not a title of the Final Project is acceptable and to find out whether the title already exists or not currently done is through control and selection of the lecturers or supervisors. Sometimes the ability of the lecturer in exercising control and selection is still constrained by having to check and find out with the memory abilities of each lecturer or supervisor that may be limited so that sometimes some titles pass the observation that causes duplicate titles.

Title duplication is a common form of plagiarism in writing final project [1], [2], [3]. As one way to overcome these problems, a system is needed to find out how much the percentage of the title of the research submitted by students with the title of the research that already exists. Data from research titles that have been available on Google Scholar, which include online journals from scientific publications [4] can be used to assist in obtaining other pre-existing titles as a reference or similar titles.

The application of web scraping with CURL (Client URLs) and simple HTML DOM parser can help to retrieve title data, as a comparison of existing research title data in google scholar [5]. Web scraping is a technique for retrieving information from a website [6], [7]. CURL is useful to transfer data to and from the server with a library and command line. CURL is useful for data retrieval methods from sites [8], [9]. Simple HTML DOM parser helps manipulate HTML elements that can work with HTML code that does not include W3C validation because Simple HTML DOM parsers are not limited to valid HTML classes. DOM elements can also be deleted, added, or changed. In HTML DOM data retrieval is based on tags, classes, IDs, and so on [10], [11].

Winnowing algorithm can be used to find the percentage of the similarity of the text of the research title proposed with the research title data from Google Scholar. Google Scholar is one of the references for search engine scientific publications so that data from the Google Scholar is a proper scientific work data used as a comparison in detecting the proposed title of the final assignment of student research.

The winnowing algorithm has fulfilled the prerequisites of the text similarity detection algorithm, namely whitespace insensitivity, i.e., only characters in the form of letters or numbers will be processed further and discard all irrelevant characters such as punctuation, spaces and other characters [12], [13]. The winnowing algorithm can detect plagiarism of text or documents even though the document has been changed in sentence structure either by spinning or paraphrasing techniques [14]. Compared to the Rabin-Karp algorithm, the winnowing algorithm produces a better percentage level with a faster processing time [15]. Previous research [16], [17], [18], [19], [20] has been carried out, but each study has not collaborated and utilized Google Scholar resources, as comparable data for the Final Project title using the Winnowing Algorithm.

Based on these problems, to reduce plagiarism and detect early submission of student research titles, a study was conducted entitled "Web Scraping and Winnowing Algorithms for Plagiarism Detection of Final Project Titles".

2. Research Methods

2.1. Related Works

Table 1 Research related to web scraping, winnowing algorithms, and google scholar include:

1. This study built a system to collect parallel corpus between Indonesian and English. The scraping process with the HTML DOM method has produced parallel corpus documents of 38,712 pairs [17].
2. This research builds a system to detect thesis titles using a winnowing algorithm to facilitate the final task coordinator or Chair of the Study Program in determining the percentage of similarities. The system in this study will detect the similarity of a title entered with the title data that has been stored in the database [18].
3. This research builds a website that is useful for finding the desired collection of journals. This website was created to streamline the search for scientific journals in the Mendeley and google scholar by utilizing ParsCit citation extraction paper data [19].
4. This study discusses the use of google scholar, which makes it easier for final level students to find legitimate reference sources for thesis assignments. Google scholar also makes it easy for trial examiners to search for words or sentences plagiarized by students who copy other people's work [20].

Table 1. Comparison of Related Research

No.	Research	Web Scraping	Winnowing Algorithm	Google Scholar
1.	[17]	Yes	No	No
2.	[18]	No	Yes	No
3.	[19]	No	No	Yes
4.	[20]	No	No	Yes
5.	Proposed Research	Yes	Yes	Yes

2.2. Web Scraping Architecture from Google Scholar

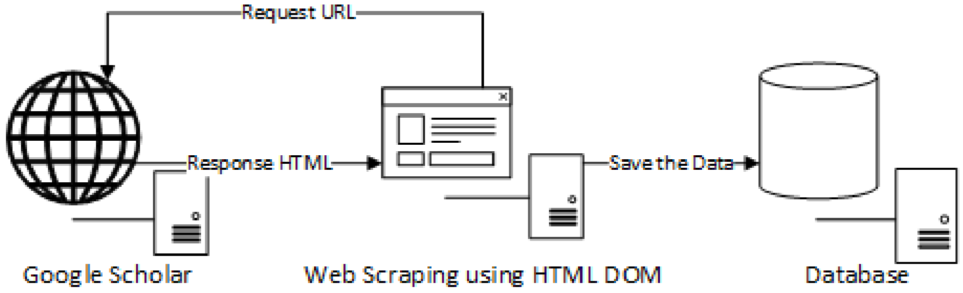


Figure 1. Web Scraping Architecture from Google Scholar

Figure 1 is a web scraping architecture. The web application requests Google Scholar, and then Google Scholar responds with HTML resources. Simple HTML DOM is used to convert HTML data and manipulate HTML elements for retrieving the data needed namely title data. Then the storage is carried out on the database, and the data is compared using a winnowing algorithm so that the comparison results with the value data in the form of a percentage of plagiarism.

2.3. Flowchart of Plagiarism Detection using Web Scraping and Winnowing Algorithms

Figure 2 a web scraping flowchart and winnowing algorithm. First, the user enters the title that will be checked by plagiarism, then the system with web scraping will retrieve the title data from the Google Scholar according to what was entered by the user. Next is the title data from Google Scholar compared to the similarity with the title entered by the user using the Winnowing algorithm. The last process of the system will display information on title data along with the percentage of similarity.

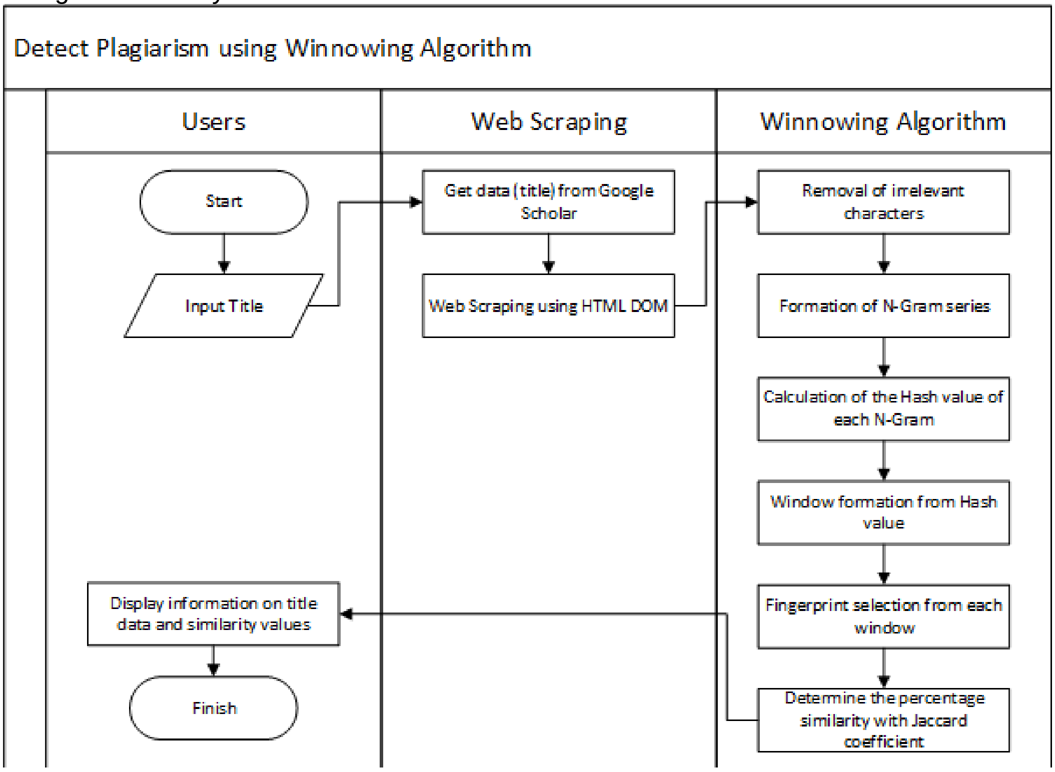

Detect Plagiarism using Winnowing Algorithm
Users
Web Scraping
Winnowing Algorithm
Start
Input Title
Get data (title) from Google Scholar
Web Scraping using HTML DOM
Removal of irrelevant characters
Formation of N-Gram series
Calculation of the Hash value of each N-Gram
Window formation from Hash value
Fingerprint selection from each window
Determine the percentage similarity with Jaccard coefficient
Display information on title data and similarity values
Finish

Figure 2. Flowchart of Plagiarism Detection using Web Scraping and Winnowing Algorithms

2.4. Textual Analysis

This system is expected to help to reduce the occurrence of duplication of research titles or plagiarism. The user checks by entering the final project title. Furthermore, the system will retrieve title data with web scraping from Google Scholar according to the title entered by the user. The title data from Google Scholar will be processed with a winnowing algorithm to find the percentage similarity between the titles entered by the user and the title of the Google Scholar.

2.5. *Use Case*

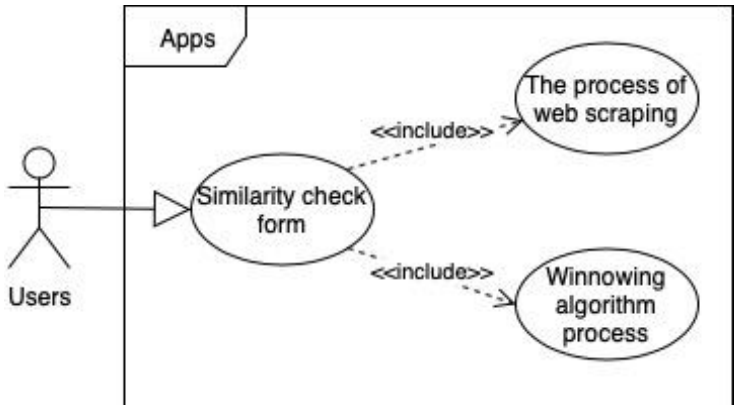
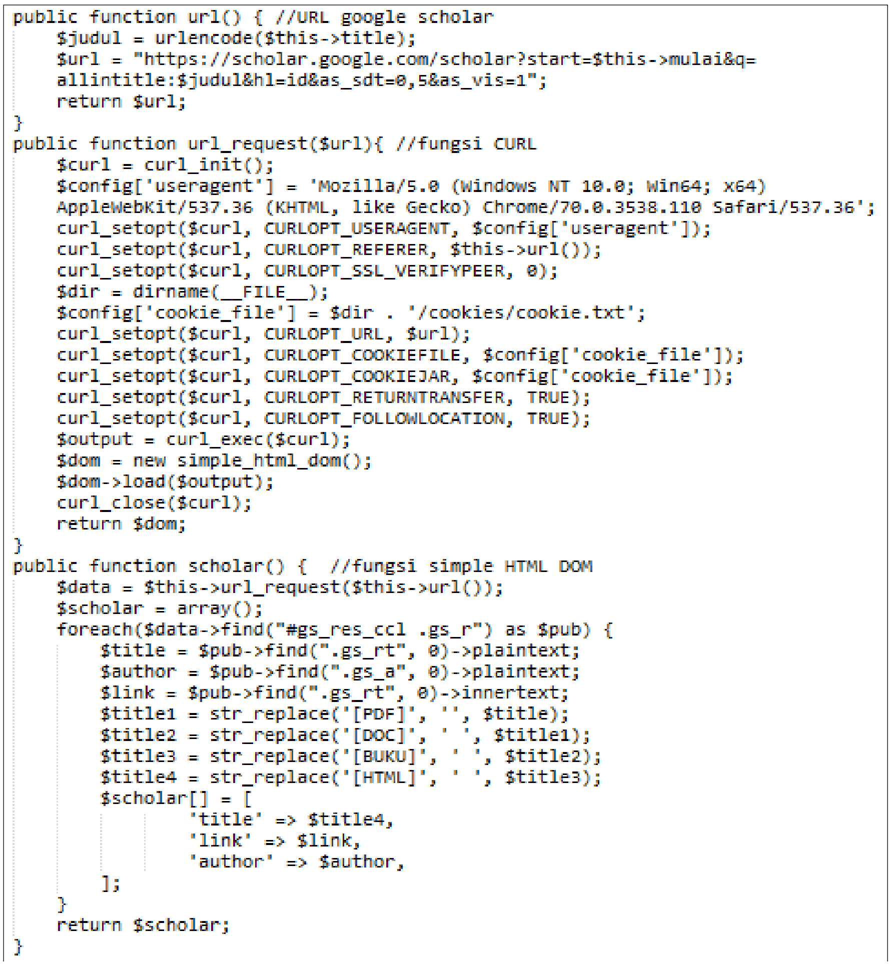


Figure 3. *Use Case* Diagram

The similarity check form in Figure 3 is a menu for checking the similarity of research titles with other research titles that already exist in Google Scholar by entering the research title to be searched for or checked for similarity. Web scraping is used to retrieve data from other research titles that already exist in Google Scholar as a reference or comparison. The process of finding the percentage similarity of the research title using the Winnowing algorithm by comparing the titles entered by the actor with the final project title data from Google Scholar.

2.6. Coding



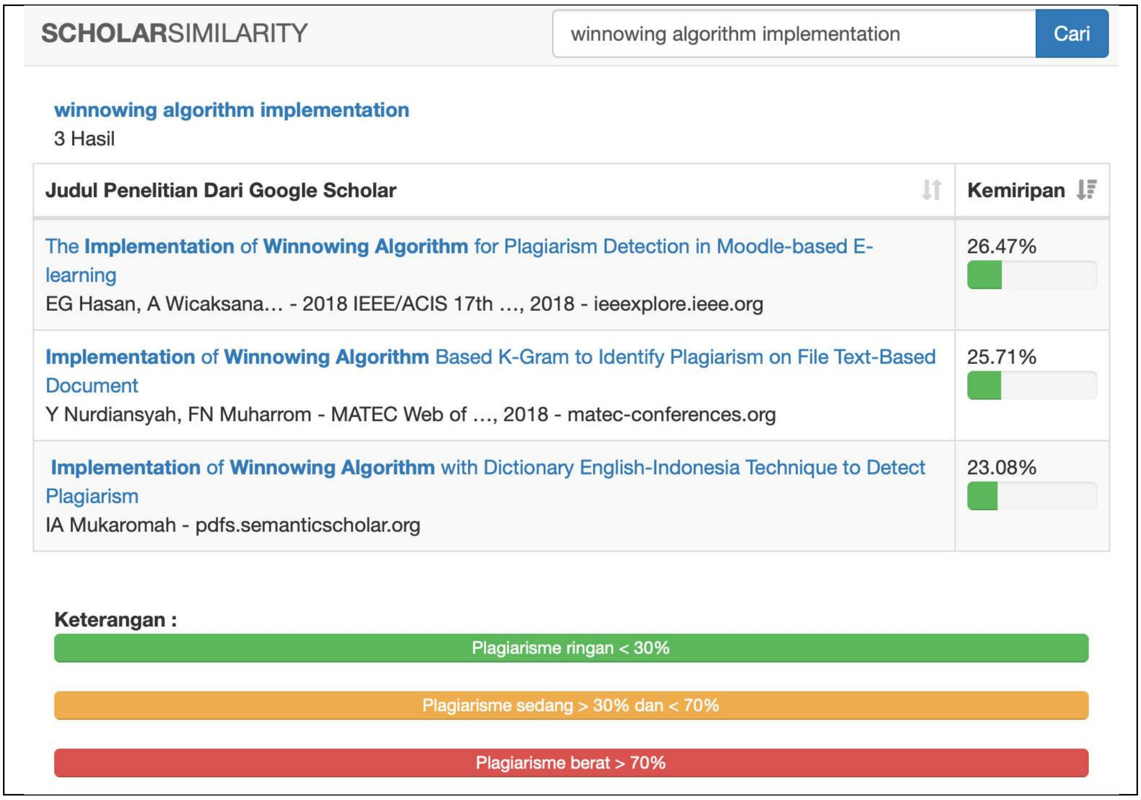
```
public function url() { //URL google scholar
    $judul = urlencode($this->title);
    $url = "https://scholar.google.com/scholar?start=$this->mulai&q=
    allintitle:$judul&hl=id&as_sdt=0,5&as_vis=1";
    return $url;
}
public function url_request($url){ //fungsi CURL
    $curl = curl_init();
    $config['useragent'] = 'Mozilla/5.0 (Windows NT 10.0; Win64; x64)
    AppleWebKit/537.36 (KHTML, like Gecko) Chrome/70.0.3538.110 Safari/537.36';
    curl_setopt($curl, CURLOPT_USERAGENT, $config['useragent']);
    curl_setopt($curl, CURLOPT_REFERER, $this->url());
    curl_setopt($curl, CURLOPT_SSL_VERIFYPEER, 0);
    $dir = dirname(__FILE__);
    $config['cookie_file'] = $dir . '/cookies/cookie.txt';
    curl_setopt($curl, CURLOPT_URL, $url);
    curl_setopt($curl, CURLOPT_COOKIEFILE, $config['cookie_file']);
    curl_setopt($curl, CURLOPT_COOKIEJAR, $config['cookie_file']);
    curl_setopt($curl, CURLOPT_RETURNTRANSFER, TRUE);
    curl_setopt($curl, CURLOPT_FOLLOWLOCATION, TRUE);
    $output = curl_exec($curl);
    $dom = new simple_html_dom();
    $dom->load($output);
    curl_close($curl);
    return $dom;
}
public function scholar() {  //fungsi simple HTML DOM
    $data = $this->url_request($this->url());
    $scholar = array();
    foreach($data->find("#gs_res_ccl .gs_r") as $pub) {
        $title = $pub->find(".gs_rt", 0)->plaintext;
        $author = $pub->find(".gs_a", 0)->plaintext;
        $link = $pub->find(".gs_rt", 0)->innertext;
        $title1 = str_replace('[PDF]', '', $title);
        $title2 = str_replace('[DOC]', ' ', $title1);
        $title3 = str_replace('[BUKU]', ' ', $title2);
        $title4 = str_replace('[HTML]', ' ', $title3);
        $scholar[] = [
                'title' => $title4,
                'link' => $link,
                'author' => $author,
        ];
    }
    return $scholar;
}
```

Figure 4. Source Code for Title Data Collection

Figure 4 is a code for web scraping programs using PHP to retrieve research title data from Google Scholar. Retrieving title data is per page with many titles, which are ten titles. Function url_request () is CURL which is used to send user agent information to Google Scholar like a web browser so that Google Scholar considers requests made by a user using a web browser and stores cookies given by Google Scholar. The function scholar () has a function to get the title data obtained by manipulating the Google scholar HTML data based on the id using the function of the simple HTML DOM parser library.

3. Result and Discussion

The user checks the similarity of the title by filling out the input form "enter the title". After filling in the title input form and pressing the search button, the system will display the research title data obtained from Google Scholar along with the percentage of similarities shown in Figure 5.



winnowing algorithm implementation

winnowing algorithm implementation
3 Hasil

Judul Penelitian Dari Google Scholar	Kemiripan
The **Implementation** of **Winnowing Algorithm** for Plagiarism Detection in Moodle-based E-learning EG Hasan, A Wicaksana… - 2018 IEEE/ACIS 17th …, 2018 - ieeexplore.ieee.org	26.47%
Implementation of **Winnowing Algorithm** Based K-Gram to Identify Plagiarism on File Text-Based Document Y Nurdiansyah, FN Muharrom - MATEC Web of …, 2018 - matec-conferences.org	25.71%
Implementation of **Winnowing Algorithm** with Dictionary English-Indonesia Technique to Detect Plagiarism IA Mukaromah - pdfs.semanticscholar.org	23.08%

Keterangan :

🟩 Plagiarisme ringan < 30%

🟨 Plagiarisme sedang > 30% dan < 70%

🟥 Plagiarisme berat > 70%

Figure 5. Display menu looking for title similarity

3.1. Black-Box Testing

Black-box testing is a method for testing software in terms of functional specifications without testing the design and program code. Testing is intended to find out whether the functions, inputs, and outputs of the software are by what is needed. Table 2 is the result of black-box testing in the application made

Table 2. Black Box Testing

Data Input	Scenario	Result
The title of the research to be sought	Will display the title data obtained from Google Scholar along with the percentage of similarity	Success
The title of the research to be searched is not available on Google Scholar	Will not display the research title data including the percentage of similarity	Success

3.2. Testing the Winnowing Algorithm manually, using the system and tools plagiarism

3.2.1. Manual Testing

The manual calculation is a calculation carried out directly by humans without using an application. The process of detecting the similarity of the first title "Implementasi Teknik *Web Scraping* Pada Aplikasi Pemesanan Tiket Kereta Api" to the second title "Implementasi Teknik *Web Scraping* Pada Aplikasi Pemesanan Tiket Pesawat".

a. Discard irrelevant characters and change all letters to lowercase in the first and second title text.

First title:

```
implementasiteknikwebscrapingpadaaplikasipemesanantiketkeretaapi
```

Second title:

implementasiteknikwebscrapingpadaaplikasipemesanantiketpesawat

b. The formation of the n-gram circuit with n = 6, it will form as follows:

n-gram first title:

implem mpleme plemen lement ementa mentas entasi ntasit tasite asitek sitekn itekni teknik eknikw knikwe nikweb ikwebs kwebsc webscr ebscra bscrap scrapi crapin raping apingp pingpa ingpad ngpada gpadaa padaap adaapl daapli aaplik aplika plikas likasi ikasip kasipe asipem sipeme ipemes pemesa emesan mesana esanan sanant ananti nantik antike ntiket tiketk iketke ketker etkere tkeret kereta eretaa retaap etaapi

n-gram second title:

implem mpleme plemen lement ementa mentas entasi ntasit tasite asitek sitekn itekni teknik eknikw knikwe nikweb ikwebs kwebsc webscr ebscra bscrap scrapi crapin raping apingp pingpa ingpad ngpada gpadaa padaap adaapl daapli aaplik aplika plikas likasi ikasip kasipe asipem sipeme ipemes pemesa emesan mesana esanan sanant ananti nantik antike ntiket tiketp iketpe ketpes etpesa tpesaw pesawa esawat

c. Calculates the hash value in the first n-gram series "impleme", base value (b) = 3, and n-gram circuit length (n) = 6.

$$H_{(\text{implem})} = ascii(i) * 3^5 + \ ascii(m) * 3^4 + ascii(p) * 3^3 + ascii(l) * 3^2 + ascii(e) * 3^1 + ascii(m)$$

$$H_{(\text{implem})} = 105 * 243 + \ 109 * 81 + \ 112 * 27 + \ 108 * 9 + \ 101 * 3 + 109$$

$$H_{(\text{implem})} = 38752$$

The results of all calculations of the first title hash value are:

38752 39812 40085 38723 37534 39088 37908 40211 40544 37175 40922 39036 40670 37565 39167 39596 38713 39693 41190 36916 37231 40356 37343 39961 36889 40051 38605 39367 38008 39049 35607 36213 35846 36922 40168 38961 38263 38345 37141 40811 38713 39691 37535 39073 37868 40091 36543 39023 36980 40343 40946 38375 38694 38180 41027 38614 37936 40291 37872

The results of all calculations of the second title hash value are:

38752 39812 40085 38723 37534 39088 37908 40211 40544 37175 40922 39036 40670 37565 39167 39596 38713 39693 41190 36916 37231 40356 37343 39961 36889 40051 38605 39367 38008 39049 35607 36213 35846 36922 40168 38961 38263 38345 37141 40811 38713 39691 37535 39073 37868 40091 36543 39023 36980 40343 40951 38390 38740 38314 41432 39829 37955

d. Setting a window with w = 4.

Window first title:

W-1 : {38752 39812 40085 38723}

W-2 : {39812 40085 38723 37534}

W-3 : {40085 38723 37534 39088}

. . .

W-56 : {38614 37936 40291 37872}

Window second title:

W-1 : {38752 39812 40085 38723}

W-2 : {39812 40085 38723 37534}

W-3 : {40085 38723 37534 39088}
. . .
W-54 : {38314 41432 39829 37955}

e. The selection of fingerprint values from the window formation.

fingerprint first title:

38723 37534 37908 37175 37565 38713 36916 37231 36889 38008 35607 35846 36922 38263 37141 37535 36543 36980 38375 38180 37936 37872

fingerprint second title:

38723 37534 37908 37175 37565 38713 36916 37231 36889 38008 35607 35846 36922 38263 37141 37535 36543 36980 38390 38314 37955

f. Jaccard coefficient:

The same fingerprint from the first title and the second title:

(38723 37534 37908 37175 37565 38713 36916 37231 36889 38008 35607 35846 36922 38263 37141 37535 36543 36980) = 18

The entire *fingerprint* is first and second title:

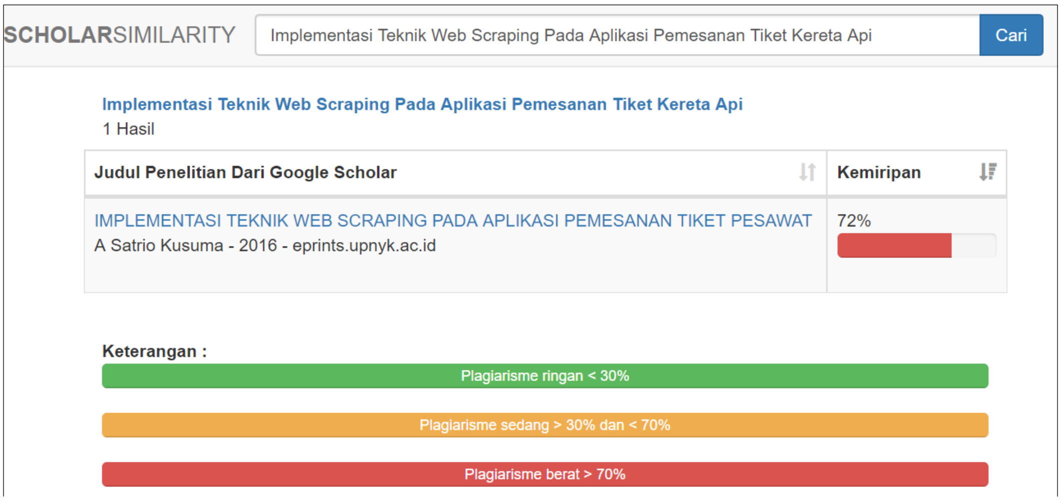
(38723 37534 37908 37175 37565 38713 36916 37231 36889 38008 35607 35846 36922 38263 37141 37535 36543 36980 38375 38390 38180 38314 37936 37955 37872) = 25

Similarity :

$$\mathit{Similarity} = \frac{18}{25} x\ 100 = 72\%$$

Percentage of text similarity between first title and second title based on the results of the similarity of the two fingerprints with a manual calculation of 72%.

3.2.2. Calculations on the system



Implementasi Teknik Web Scraping Pada Aplikasi Pemesanan Tiket Kereta Api

Implementasi Teknik Web Scraping Pada Aplikasi Pemesanan Tiket Kereta Api
1 Hasil

Judul Penelitian Dari Google Scholar	Kemiripan
IMPLEMENTASI TEKNIK WEB SCRAPING PADA APLIKASI PEMESANAN TIKET PESAWAT A Satrio Kusuma - 2016 - eprints.upnyk.ac.id	72%

Keterangan :
Plagiarisme ringan < 30%
Plagiarisme sedang > 30% dan < 70%
Plagiarisme berat > 70%

Figure 6. The results of the calculation of the winnowing algorithm on the system

Figure 6 shows the results of the calculation of the system winnowing algorithm with a value of n = 6, w = 4, and b = 3, with the results of 72% similarity.

These results indicate that the calculation of the manual winnowing algorithm and the system get the same results, namely 72%. Plagiarism can be grouped according to proportion or percentage of sentences or hijacked paragraphs, namely mild plagiarism (<30%), moderate plagiarism (30–70%) and severe plagiarism (> 70%) [21] [22].

3.3. Testing with Plagiarism Checker X Tools

This test was conducted to compare the results of the percentage similarity between the systems proposed in this study with tools plagiarism checker X.Plagiarism Checker X is a tool to help detect plagiarism in research papers, blogs, assignments, and websites. To find the percentage of the title similarity to the X checker plagiarism application is done by side by side comparisons by entering the tested title and the comparison title.

Table 3. The title tested and the comparison title

No	Tested Title	Comparative Title
1.	Implementation of Web Scraping Techniques on Train Ticket Booking Applications	Implementation of Web Scraping Techniques in Airplane Ticket Booking Applications
2.	Implementation of RESTful Web Service for Election Vote Calculation System	Implementation of RESTful Web Service for Rapid Vote Counting System in Local Election
3.	CRM Implementation to Increase Customer Loyalty	Analysis of Electronic CRM Implementation at PT Cordova Garment to Increase Customer Loyalty
4.	Medical Record Information System at RSUD Pacitan General Hospital Based on Android	Medical Record Information System at the Regional General Hospital of RSUD Pacitan Based on Web Base
5.	Similarity Thesis Detection System using Rabin Karp's Algorithm	Thesis Title Similarity Detection System Using Winnowing Algorithms
6.	Scientific Article Search Website by Utilizing Google Scholar and Mendeley API	Website Search for Scientific Articles by Utilizing Parscit's Google Scholar and Mendeley API
7.	Web Scraping Implementation on Ontology-Based Web for Drug Data	Web Scraping Implementation on Ontology-Based Web for Drug Data and Disease
8.	Implementation of Customer Relationship Management in the Hotel Reservation System	Implementation of Customer Relationship Management CRM in a Website and Desktop-based Hotel Reservation System
9.	Designing Information Systems for competitive advantages of modern companies	Analysis and Design of Information Systems for competitive advantages of modern companies and organizations
10.	Information System Distribution of Information Technology Research Sites in Garut	Designing Geographic Information Systems Distribution of Information Technology Research Sites in the City of Garut
11.	Designing Achievement Decision Selection System for Student Achievement	Designing the Decision Support System for the Selection of Outstanding Students using the AHP and Promethee Methods

Table 3 is the title data tested and the title data as a comparison so that the percentage value of plagiarism will be obtained using the system proposed in the study with tools plagiarism checker X.

Table 4. Similarity percentage comparison

No.	This Research	Plagiarism Checker X tools
1.	72%	89%
2.	68.75%	67%
3.	38.89%	0%
4.	80.65%	86%
5.	70.37%	88%
6.	87.5%	92%
7.	83.87%	85%
8.	58.97%	62%
9.	54.29%	58%

10.	46.47%	46%
11.	67.57%	62%
Average	**66.30%**	**66.82%**

Table 4 is the percentage data of the plagiarism value from the comparison between the systems proposed in the study with tools plagiarism checker X. The system created has a smaller percentage average of 66.30% compared to X plagiarism checker application, with an average of 66.82%.

4. Conclusion

Based on the results of testing in the study conclusions can be drawn, namely; Web scraping with CURL and simple HTML DOM parser can be applied to retrieve data from Google Scholar's research title on early detection applications for submitting student research titles. Google Scholar can be used to obtain other existing research titles as a reference or comparison in early detection applications submitting student research titles by applying web scraping as a method of retrieving data. Winnowing algorithm can be applied to find the percentage similarity of the research title proposed with the existing research title in Google Scholar in the application of early detection submission of student research titles. This research is still lacking. Namely, the comparative title data source only from Google Scholar and the data compared only to the title, can not know the author of the scientific work. Also, the application of the method in this study has not been able to detect research titles with different languages.

References

[1] N. Knock dan R. Davison, “Dealing with Plagiarism in the Information Systems,” *MIS Quarterly,* vol. 27, pp. 511-532, 2003.

[2] Mulyana, “Pencegahan Tindak Plagiarisme Dalam Penulisan Skripsi,” *Cakrawala Pendidikan,* 2010.

[3] A. Y. Gasparyan, B. Nurmashev, B. Seksenbayev, V. I. Trukhachev, E. I. Kostyukova dan G. D. Kitas, “Plagiarism in the Context of Education and Evolving Detection Strategies,” *Journal of Korean Medical Science,* vol. 32, no. 8, pp. 1220-1227, 2017.

[4] Google, “Tentang Google Cendikia,” [Online]. Available: https://scholar.google.com/intl/id/scholar/ about.html. [Diakses 9 September 2018].

[5] R. Gunawan, A. Rahmatulloh, I. Darmawan dan F. Firdaus, “Comparison of Web Scraping Techniques: Regular Expression, HTML DOM and Xpath,” dalam *2018 International Conference on Industrial Enterprise and System Engineering (IcoIESE 2018)*, Atlantis Press, 2019.

[6] B. G. Dastidar, D. Banerjee dan S. Sengupta, “An Intelligent Survey of Personalized Information Retrieval using Web Scraper,” *International Journal of Education and Management Engineering,* vol. 5, no. 3, pp. 24-31, 2016.

[7] M. Turland, “php| architect's Guide to Web Scraping with PHP,” Marco Tab ini&Associates, 2010.

[8] D. Stenberg, “CURL: curl groks URLs,” 2015.

[9] M. I. Khalid, PHP/CURL Book with Examples Version 1.8, 2006.

[10] V. B. Kadam dan G. K. Pakle, “A Survey on HTML Structure Aware and Tree Based Web Data Scraping Technique,” *International Journal of Computer Science and Information Technologies (IJCSIT),* vol. 5, no. 2, pp. 1655-1658, 2014.

[11] V. Janjic, “PHP Simple HTML DOM Parser: Editing HTML Elements in PHP,” 7 September 2011. [Online]. Available: https://phpbuilder.com/php-simple-html-dom-parser-editing-html-elements-in-php/. [Diakses 6 Oktober 2018].

[12] X. Duan, M. Wang dan J. Mu, “A Plagiarism Detection Algorithm based on Extended Winnowing,” dalam *2017 International Conference on Electronic Information Technology and Computer Engineering (EITCE 2017)*, 2017.

[13] S. Schleimer, D. S. Wilkerson dan A. Aiken, “Winnowing: Local Algorithms for Document Fingerprinting,” *Proceedings of the ACM SIGMOD international conference on management of data,* pp. 76-85, 2003.

[14] H. Tri Nugroho I, “Pengaruh Algoritma Stemming Nazief-Adriani Terhadap Kinerja Algoritma Winnowing Untuk Mendeteksi Plagiarisme Bahasa Indonesia,” *ULTIMA Computing,* vol.9, no. 1, pp. 36-40, 2017.

[15] N. Alamsyah, “Perbandingan Algoritma Winnowing dengan Algoritma Rabin Karp untuk Mendeteksi Plagiarisme pada Kemiripan Teks Judul Skripsi,” *Technologia,* vol. 8, no. 3, pp. 124-134, 2017.

[16] I. P. A. Darmawan dan I. N. P. I. P. A. Dharmaadi, “Ekstrak Hirarki Data Dari Situs Web A-Z Animals Menggunakan Web Scraping,” *Lontar Komputer : Jurnal Ilmiah Teknologi Informasi,* vol. 8, no. 3, pp. 124-134, 2017.

[17] V. Mitra, H. Sujaini dan A. B. Putra Negara, “Rancang Bangun Aplikasi Web Scraping Untuk Korpus Paralel Indonesia - Inggris Dengan Metode HTML DOM,” *Jurnal Sistem dan Teknologi Informasi (JUSTIN),* vol. 5, no. 1, pp. 36-41, 2017.

[18] Nurdin dan A. Munthoha, “Sistem Pendeteksi Kemiripan Judul Skripsi Menggunakan Algoritma Winnowing,” *InfoTekJar (Jurnal Nasional Informatika dan Teknologi Jaringan),* vol. 2, no. 1, pp. 90-97, 2017.

[19] I. Ruslan, A. Wibowo dan R. Lim, “Website Penelusuran Artikel Ilmiah dengan Memanfaatkan Parscit, Google Scholar, dan Mendeley Api,” *Jurnal Infra, vol. 1, no. 2,* 2013.

[20] K. Tiara, U. Rahardja dan I. A. Rosalinda, “Pemanfaatan Google Scholar Dan Citation Dalam Memenuhi Kebutuhan Pembuatan Skripsi Mahasiswa Pada Perguruan Tinggi,” *Technomedia Journal (TMJ),* vol. 1, no. 1, pp.95113, 2016.

[21] S. Sastroasmoro, “Beberapa Catatan tentang Plagiarisme,” Majalah Kedokteran Indonesia, vol. 57, no. 8, Agustus, 2007.

[22] J. D. Velásquez dan E. M. Taylor, “Tools for External Plagiarism Detection in DOCODE,” dalam *WI-IAT '14 Proceedings of the 2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)*, 2014.