

Building A Web-Based Final Project Detection Information System with Incremental Method and Jaccard Similarity Algorithm

Maria Eryestin Nahak¹, Darsono Nababan², Yasinta O.L Rema³

^{1,2,3}Information of Technology, Faculty of Agriculture, University of Timor
^{1,2,3}Jl. Km. 09 Kelurahan Sasi, Kecamatan Kota Kefamenanu, Kabupaten TTU, Provinsi Nusa Tenggara Timur, Indonesia.

ABSTRACT

Article:

Accepted: April 16, 2023

Revised: March 11, 2023

Issued: April 30, 2023

© 2023 The Author(s).



This is an open-access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license

*Correspondence Address:

eryestinnahak@gmail.com

Thesis is a scientific work and one of the graduation requirements for students in tertiary institutions. The process of implementing the thesis at the Information Technology Study Program at the University of Timor is still done manually, starting from submitting titles to data processing which allows for duplicate titles, difficulty finding data and losing data. For this reason, a system is needed to avoid the risk of duplicate titles and difficulty finding the data. So in this study an information system will be built to detect the similarity of the title of the final project (Thesis) using the incremental method and the Jaccard similarity algorithm as a plagiarism detection algorithm. The application of the incremental method is due to its flexible nature and less risk of failure than other development methods. System testing is carried out on the functional system and using the User Acceptance Test. Based on the test results, this system facilitates Study Programs in submitting final assignment titles, finding final assignment data, minimizing the risk of duplicate titles through the Title Similarity Percentage testing feature to find out the highest level of similarity based on thesis title category and avoiding data loss.

Keywords: *information system, plagiarism, thesis, incremental, jaccard similarity.*

1. INTRODUCTION

The rapid development of increasingly modern technology has aspects of influence in various fields in increasing the competence of human resources. One of them is the world of education. Based on the Law of the Republic of Indonesia No. 12 of 2012 concerning Higher Education which sets out the Tridharma of Higher Education and the Obligations of Higher Education to organize education, research and community service [1]. Implementation of Tridharma is a form of student responsibility to conduct research as a final assignment which is a requirement for graduation from the Bachelor's program.

The Final Project is a scientific paper, in the form of an explanation of research results to discuss factual problems by applying scientific principles that apply to solving certain problems [2]. One of the applied scientific methods is the incremental method. The application of the incremental method is because it is flexible and the risk of failure is minimal compared to other development methods [3]. The incremental method is wrong one method of software development which can minimize the loss adjustments in the development process software . On method increment, every stage that is in methodology there is input (input) and output (output). Process output increment will be used as input (input) for the next increment process [4].

Previously, there was previous research that discussed how to implement the Jaccard similarity algorithm in checking the similarity of thesis titles at Department of Computer Engineering Sriwijaya State Polytechnic, which discussed the application of the Jaccard similarity algorithm to see the percentage level of similarity in the title of the final project report. In his research, the researcher built a plagiarism detection system for report titles to determine the percentage level of similarity with existing titles. Researchers are aware of the shortcomings in their research, namely in terms of developing methods and appearance of the detection system [5].

Meanwhile, the registration of thesis titles and thesis guidance at the Information Technology Study Program at the University of Timor is still done manually where students submit thesis titles by bringing the draft title to the head of the study program at a certain place and time. In addition, the search for the wrong

time in the title submission process can lead to time ineffectiveness. Likewise, the data recap of the implementation of the final project which is recapitulated manually one by one and stored in folders made by the lecturer or the head of the study program. This allows for duplicate titles, where titles that have been submitted can be submitted again by other students. Not only that, the limitations of the current system cause difficulties in finding the work of the final project, due to limited search methods.

So it is necessary to have a system that is able to detect similarities in titles, overcome duplicate titles, and facilitate the search for student final assignments in the previous period. This system is also capable of processing student final assignment data such as digital storage media for scientific work, facilitating data search, and can display final project work and is able to provide important information regarding the percentage of students doing final assignments and the percentage of student titles based on topic categories title in the Information Technology Study Program at the University of Timor.

2. METHODS

This type of research is applied research. This type is directed to innovate in the development of science and technology with the aim of getting solutions to problems in society, industry and government as a continuation of pure research [6]

The method used in research for the development of the SCRIPTI system is the incremental method, which is a type of system model development that focuses on adding features at each iteration. The incremental method aims to overcome the weaknesses of the waterfall method which does not accommodate iterations, and the weakness of the prototype method which has a process that is too short and each iteration of the process does not always produce a product (it could be just a prototype). The incremental incremental method is accommodated to produce features at each incremental stage [7]

The following is an illustration of the Incremental method.

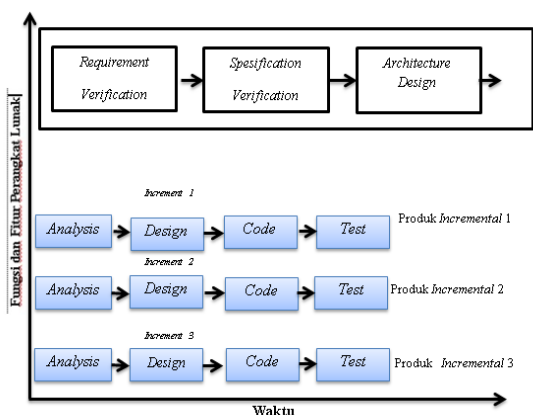


Figure 1 Illustration of the Incremental Method

The coherent stages of the incremental method are described in the following flowchart:

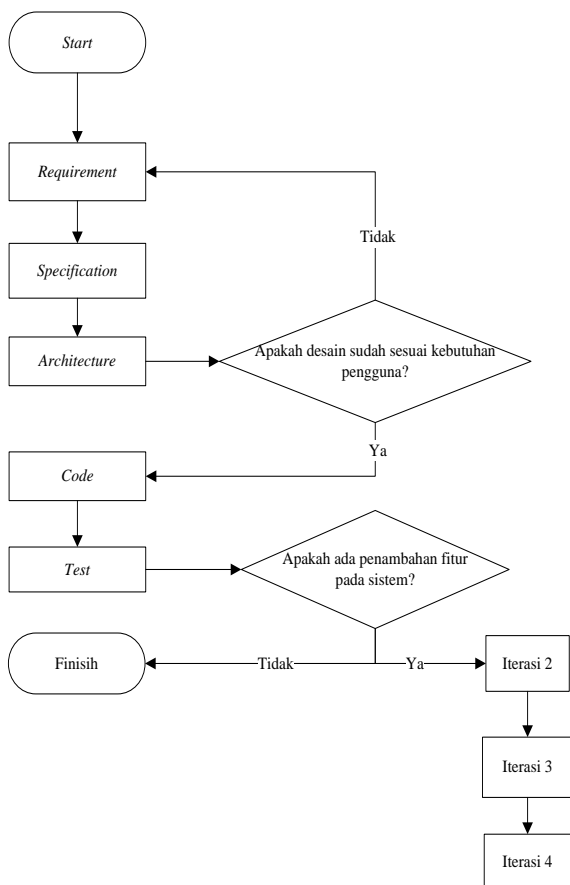


Figure 2 Incremental Method Flow

The following are the stages of system development by applying the incremental method.

a. Requirements

At the requirements stage, the researcher conducts a needs analysis for the system to be built by collecting data related to the system [8].

b. Specifications

The process of detailing user requirements for existing systems at the requirements stage. The goal is that the flow of the system and user needs can be identified clearly.

c. Architecture

This process includes system modeling using use cases and activity diagrams. There is a database design and system mockup design according to system requirements[9]. When the architecture stages are completed, it will be decided whether the system design is in accordance with the requirements? If appropriate, proceed to the coding stage. However, if it is not appropriate, then the needs analysis stage is repeated.

d. Code

Purpose of Coding to build an information system for detecting similarity of the title of the final project (Thesis) using HTML and PHP syntax and a MySQL database, as well as implementing the Jaccard similarity algorithm for the thesis title similarity detection feature

e. Test

Testing is carried out by testing the success of features to find out how far the system has been built can be used and whether the system built meets user needs or not? Or will it be incremented to add other features to the system? If there are no additional features, the process ends. However, if there are still missing features and increments will be carried out, the system development process will continue to iteration 2, and so on

3. RESULTS AND DISCUSSION

3.1. Requirements Process

This process is carried out by analyzing the title submission process and processing the final project data manually.

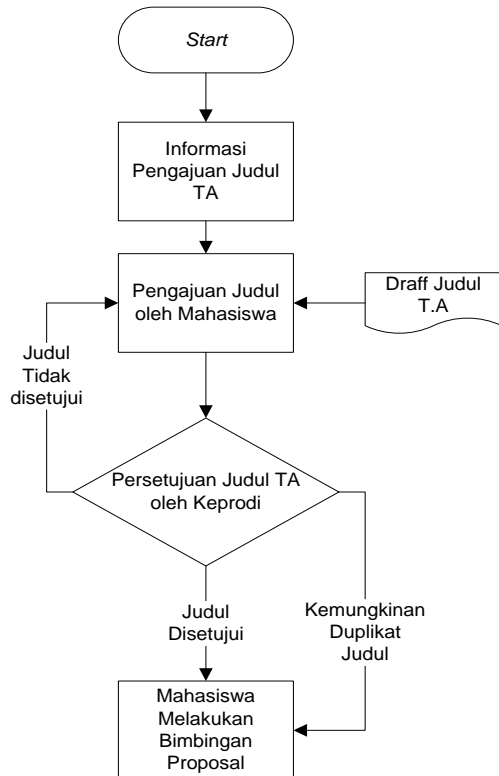


Figure 3. Old System Flowchart

The following is the proposed plagiarism system flow and named the SCRIPTI system.

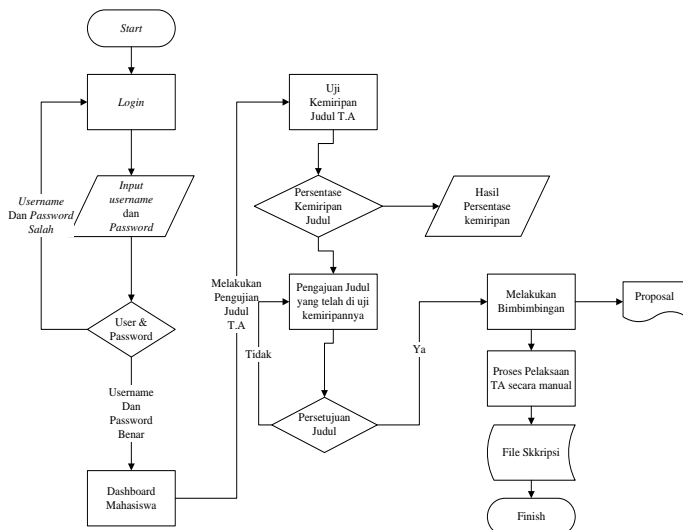


Figure 4. SCRIPTI system flowchart

3.2. Process Specification

Specifications are carried out using the following system use cases to explain how the user treats the system.

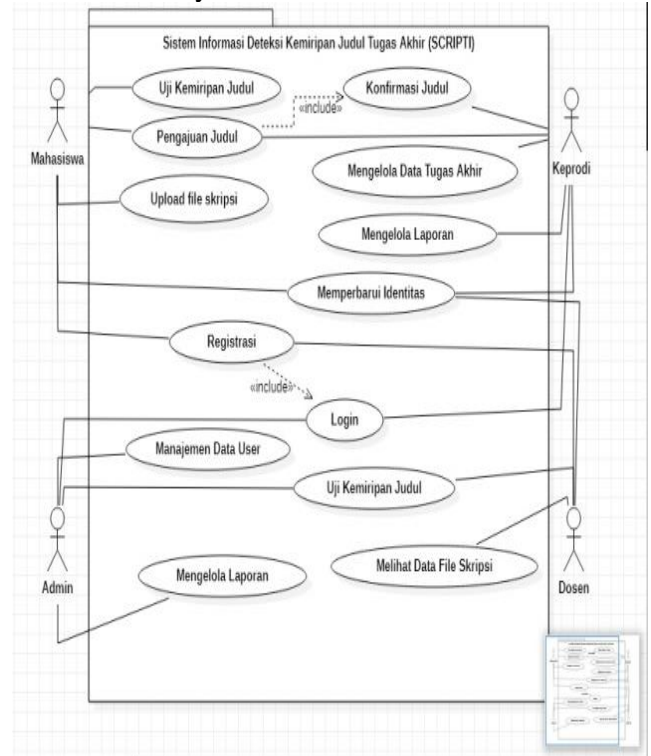


Figure 5. SCRIPTI system use case

[10] There are 2 analysis of system requirements, namely analysis of functional requirements and analysis of non-functional requirements as follows:

a. Functional Requirements Analysis

Requirements that contain any processes or services that must be provided by the system, also includes how the system must react to certain inputs and how the system behaves in certain situations such as Login Features, Dashboard Features, Title Similarity Test Features, Title Submission Features, Repository Features, Reports and Data Management Like, Add data, Delete data, Print and Edit.

b. Non-functional Needs Analysis

Non-functional requirements are requirements that focus on the behavioral properties of the system. Functional requirements are also often referred to as service or function boundaries offered by the system such as time constraints, process development limitations, standardization and others. Examples of Functional Requirements in the SCRIPTI system are

1. Usability
 which refers to the appearance of an effective and easy-to-use system,
2. Portability
 The system (SCRIPTI) is a web-based system, which can be accessed anywhere and anytime. This system also supports various web browsers on PC or labtop.
3. Reliability
 When accessing the system, the system will identify the user through the login form which contains a password that has been determined by the user and is only known by the user himself to access the system.

Specification process is also carried out using activity diagrams to find out how students use the plagiarism system

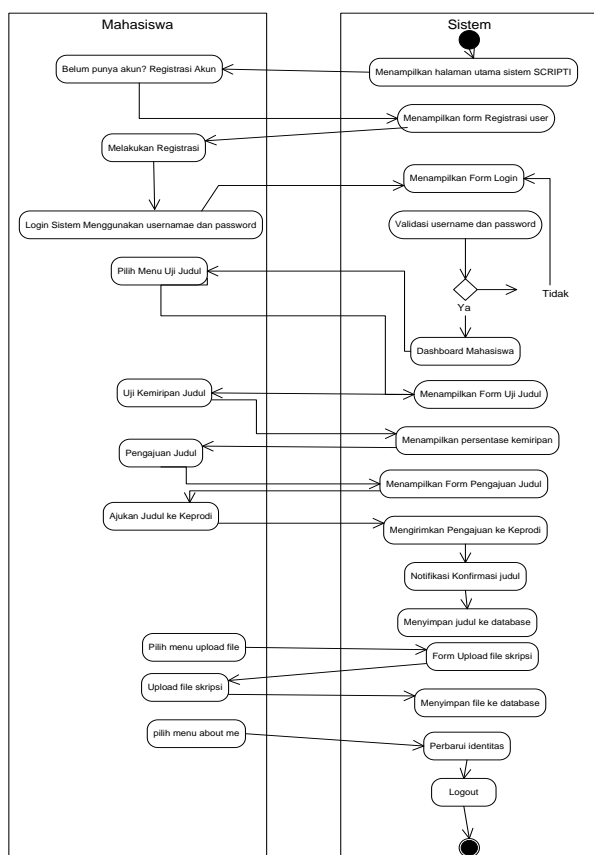


Figure 6 Diagram Activity.

3.3. Coding and System Testing

The development of the SCRIPTI system applied the Jaccard Similarity algorithm as an algorithm to detect the similarity of the title of the final project.

Jaccard similarity or jaccard coefficient is an algorithm whose function is to compare two samples, namely one document or sentence with another based on the word it has. Jaccard similarity is usually used to compare documents and calculate the similarity value of the two objects [11].

Jaccard similarity can be formulated as follows:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|} \quad (1)$$

Information :

- J = Jaccard
- A = Document 1
- B = Document 2
- $|A \cap B|$ = The number of words is the same between A and B
- $|A \cup B|$ = The total number of words that are the same and words that are different

In calculating the percentage of similarity, there are 7 fungsi PHP stages that need to be carried out sequentially in writing program code, including:

1. strtolower
2. explored
3. array_intersect
4. count
5. array_merge
6. array_unique
7. count

In Coding, the author uses PHP and HTML Programming Language as the system language. The following is the Program Code used to calculate the Percentage of Similarity to Thesis Titles.

```

<?php
session_start();
$no = 1;
if(isset($_GET['proses']))
    //mengubah semua inputan dan data dalam
    database menjadi huruf kecil

    $judul = trim($_GET['judul']);
    $kalimat_new = strtolower($judul);
echo "<center/>";echo "JUDUL TUGAS AKHIR
ANDA YANG AKAN DILAKUKAN UJI
KEMIRIPAN ADALAH :";
echo "<center/>";echo $kalimat_new;
?><hr><hr>
<?php
$koneksi =
mysqli_connect("localhost","root","","scripti");

    $no = 1;
    $sql = "SELECT * FROM data_judul ";
    $query = mysqli_query($koneksi,$sql);
    while ($data=mysqli_fetch_array($query))
    { ?>
<tr>
    <td><?php echo $no;?></td>
    <td><?php echo $data['npm'];?></td>
    <td><?php echo
        $data['nama_mahasiswa'];?></td>
    <td><?php echo $data['kategori'];?></td>
    <td><?php echo $kalimat_new1 =
        strtolower($data['judul']);?></td>
    <td><?php
        $data1 = explode(" ", $kalimat_new);
        $data2 = explode(" ",
$kalimat_new1);

        $result = array_intersect($data1, $data2);
        $result2 = count($result);

        $array1 = array_merge($data1, $data2);
        $array2 = array_unique($array1);
        $array3 = count($array2);
        $array4 = number_format($result2/$array3
            * 100,2); \
            print_r($array4);echo("%");

        $no++; } ?>
    </td>
</tr>
    
```

Text Box 1. Title Testing Feature program code

Based on manual calculations performed. The same test is also carried out on the Functional System using Black Box Testing and the system can calculate the percentage of similarity as expected

The following are the results of tests carried out to test similarity of a title "Implementation of the Incremental Method in Building a Final Title Similarity Detection Information System (Thesis) Using Web-Based Jaccard Similarity." Against the 10 final title assignments used as test samples. The test results are as follows.

Table 1. Results of the similarity test for 10 thesis titles

No	Thesis Title	Similarity
1	implementation of the incremental method in building a similarity detection information system for the title of the final project (thesis) using web-based jaccard	100%
2	implementation of the incremental method in building a web-based alumni study tracking information system	42.86%
3	the application of the waterfall method in the development of an information system for the church of Christ the king, eban	16.00%
4	geographic information systems for mapping tourist objects in TTS using the waterfall method	16.00%
5	Implementation of the importance performance analysis method in measuring the level of community satisfaction with service performance at dukcapil	10.00%
6	expert systems for diagnosing dengue fever using the Naive Bayes method	7.69%
7	implementation of Bayes' theorem in an expert system for diagnosing diseases in tomato plants	7.41%
8	classifications of woven fabrics using artificial neural networks	4.17%
9	expert systems for diagnosing malnutrition in toddlers	4.10%
10	hardware introduction android applications for elementary school students using augmented reality	3.70%

The following are stages of Jaccard Similarity Algorithm process in determining the percentage of title similarity:

1. Determine the two final project titles that will be tested for their level of similarity:
 - A. Implementation of the Incremental Method in Building a Similarity Detection Information System for Final Project Titles (Thesis) Using Web-Based Jaccard Similarity
 - B. Implementation of the Incremental Method in Building Web-Based Alumni Study Tracher Information Systems
2. Change the title sentence that is inputted and the title in the database to lowercase using "strtolower".
 - A. implementation of the incremental method in building a similarity detection information system for the title of the final project (Thesis) using web-based jaccard similarity
 - B. implementation of the incremental method in building a web-based alumni study tracker information system
3. Separate word for word in sentences "A and B" using "explode"
 - A. Array ((0) => implementation (1) => method (2) => incremental (3) => in (4) => build (5) => system (6) => information (7) => detect (8) => similarity (9) => title (10) => assignment (11) => final (12) => thesis (13) => using (14) => jaccard (15) => similarity (16) => based (17) => web)
 - B. Array ((0) => implementation (1) => method (2) => incremental (3) => in (4) => build (5) => system (6) => information (7) => tracher (8) => study (9) => alumni (10) => based (11) => web)
4. Compare word for word the same from sentences "A" and "B" with "array_intersect" and count the number of words compared using "count" :
 Array ((0) => implementation (1) => method (2) => incremental (3) => in (4) => build (5) => system (6) => information (14) => based (15) => web).

So that the total number of words is the same and similar in "A" and "B" is $|A \cap B| = 9$ words.

5. Counting the total number of words per word as a whole from sentences "A" and "B" with "array_merge" then :
 Array ((0) => implementation (1) => method (2) => incremental (3) => in (4) => build (5) => system (6) => information (7) => detect (8) => similarity (9) => title (10) => assignment (11) => final (12) => THESIS(13) => use (14) => jaccard (15) => similarity (16) => based (17) => web (18) => implementation (19) => method (20) => incremental (21) => in (22) => build (23) => system (24) = > information (25) => tracher (26) => study (27) => alumni (28) => based (29) => web)

6. Combine the same words into one by using "array_unique" then:
 Array ((0) => implementation (1) => method (2) => incremental (3) => in (4) => build (5) => system (6) => information (7) => detect (8) => similarity (9) => title (10) => assignment (11) => final (12) => thesis (13) => use (14) => jaccard (15) => similarity (16) => based (17) => web (18) => tracher (19) => study (20) => alumni (21))

7. Count the number of words that have been combined using "count" so that $|A \cup B| = 21$ words
8. Calculating the percentage of title similarity with the Jaccard Similarity algorithm

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \quad (2)$$

$$= \frac{|A \cap B|}{|A| + |B| - |A \cap B|} \quad (3)$$

$$J(A,B) = \frac{|9|}{|21|} = \frac{|9|}{|9| + |21| - |9|}$$

$$= 0,42857 \times 100\%$$

$$= 42,86\%.$$

The results of the similarity of the two titles using the Jaccard Similarity algorithm is 42.86%.

[12]determining the similarity between the input document and the test document when carrying out the similarity test, there are 5 categories of similarity percentages, namely:

1. 0%: The result of 0% similarity means that the input sentence and the test sentence in the database do not have the slightest title in common.
2. < 15%: The percentage of similarity of 1-14% means that the input sentences and test sentences in the database have little resemblance.
3. 15-50%: The percentage of similarity of 15-50% indicates that the input sentences and test sentences in the database include moderate plagiarism.
4. >50%: The percentage of similarity is more than 51-99%, meaning that the title submitted is close to plagiarism from an existing title.
5. 100% : Similarity percentage of 100% means that the title is exactly the same and is in the category of maximum plagiarism.

In addition to the Similarity Detection Feature, the scripti system also provides information regarding the percentage of data related to the size of the title category chosen by students which is presented in the form of a pie chart.

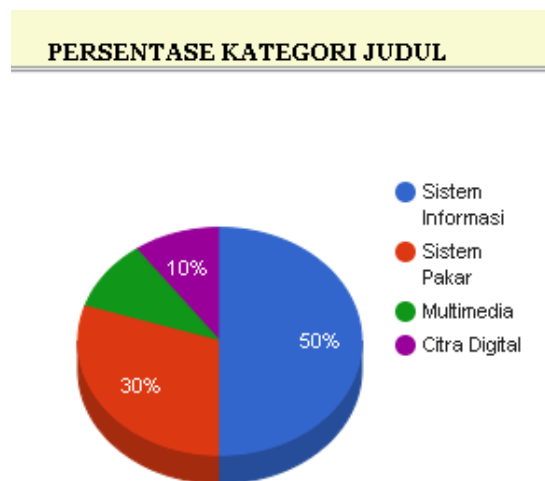


Figure 7. Diagram Title Categorial

From the diagram it is known that of the 10 titles worked on by students, students have an interest in working on information systems

topics, the percentage of which is 50%, 30% expert systems, 10% multimedia and 10% digital imagery. This can be the foundation and make it easier for study programs to made the decision to be able to increase research on other topics such as Data mining and Artificial Intelligence.

3.4. System Implementation

The resulting application is a website-based final report title similarity checking system. Using Microsoft Windows 10 as the operating system, the Sublime Text application as a text editor for creating structured programming, and the XAMPP V.3.2.2 application as a stand-alone server (localhost) which consists of Apache as a web server and MySQL as a database.

1. Student Dashboard

student dashboard will display information on how students use the system, displaying widgets for the number of titles per category of titles in the study program.

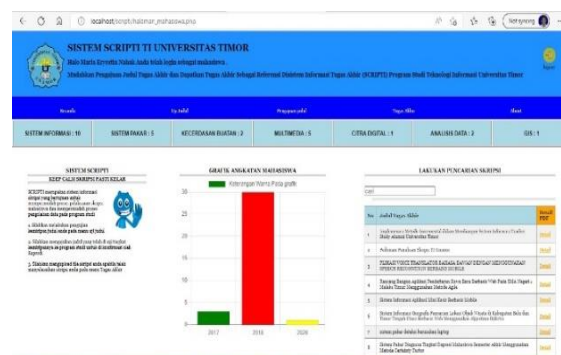


Figure 8. Student Dashboard

2. Final Project Title Similarity Test Page

This page is used to test the percentage of similarity of thesis titles submitted to study programs. In the picture below, the title test is carried out on 10 titles in the database and have different percentage results.

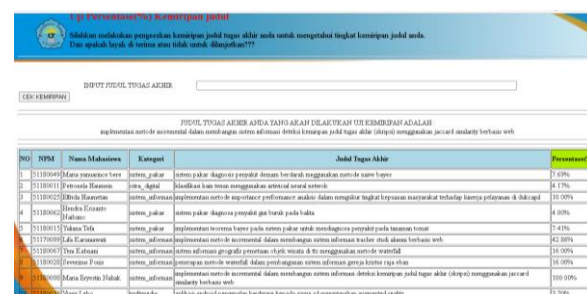


Figure 9. Display of the Title Similarity Test menu

3. Final Assignment Title Submission Page

This page is used by final project students to submit a final project title to the study program to be approved and accepted as the title of a research proposal.



Figure 10. Display of the title submission menu

4. Repositories Feature

In the SCRIPTI system there is a repository feature for student final assignments which can be used as a reference by students to read and find out what the writing structure is like and what will be the subject matter of the system in the thesis.



Figure 11. Display of Repository Features

CONCLUSION

Successfully built a plagiarism system, namely the SCRIPTI system with various features that provide data related to student final assignments as well as a means of information on title similarities when testing titles to be submitted, making it easier to find student final assignment data at the Information Technology Study Program at the University of Timor. The application of the Incremental Method to the development of this system through the stages of requirements, specification, architecture, code, test is considered effective enough to accommodate iterations to produce each new feature in its development. While the use of the Jaccard similarity algorithm that is applied as a

whole is able to detect similarities with the percentage results that are in accordance with the circumstances.

This research can be a reference in future research both from the algorithm and the system development in the future. Finally, the author also realizes the shortcomings of this research and requires suggestions for the development of this research in the future. because Building a title similarity checking application is the same as finding information that fits the user's needs, because this application is an information keyword for displaying documents which is called a system for retrieving information that has already been used.

REFERENCE

- [1] B. E. Kusuma, "Sistem Informasi Bimbingan Skripsi Berbasis Web Di Universitas Pelita Hara[1] M. Syarif and W. Nugraha, 'Metode Incremental Dalam Membangun Aplikasi Identifikasi Gaya Belajar Untuk Meningkatkan Hasil Belajar Siswa,' Jusikom J. Sist. Komput. Musirawas, vol.," *J. ISD*, vol. 3, no. 1, pp. 71–78, 2018.
- [2] H. Kurniawan and W. B. Bodowoso, "Sistem Informasi Terintegrasi Tugas Akhir/Skripsi Berbasis Web (Studi Kasus: Jurusan Sistem Informasi Institut Informatika dan Bisnis Darmajaya)," *J. Sist. Inf. dan Manaj. Basis Data*, vol. 2, no. 2, pp. 124–134, 2019.
- [3] J. Handoyo, "TUGAS AKHIR BERBASIS WEB (Studi Kasus : Sekolah Tinggi Teknologi Ronggolawe Cepu)," vol. 11, no. 1, pp. 6–13, 2017.
- [4] M. Syarif and W. Nugraha, "Metode Incremental Dalam Membangun Aplikasi Identifikasi Gaya Belajar Untuk Meningkatkan Hasil Belajar Siswa," *Jusikom J. Sist. Komput. Musirawas*, vol. 4, no. 1, pp. 42–49, 2019, doi: 10.32767/jusikom.v4i1.441.
- [5] A. Rasyid, A. Firdaus, and K. Kunci, "Aplikasi Pengecek Kemiripan Judul Laporan Akhir Berbasis Website," vol. 1, no. 1, pp. 18–27, 2021.
- [6] A. B. A. B. I. Pendahuluan, "PROGRAM STUDI TEKNOLOGI INFORMASI PROGRAM STUDI TEKNOLOGI INFORMASI," 2019.

- [7] W. Novianti, R. Amalia, F. S. Hasanusi, P. Studi, T. Informatika, and F. Teknik, "IMPLEMENTASI METODE ITERATIVE INCREMENTAL PADA SISTEM ADMINISTRASI ORGANISASI GERAKAN ANTASARI SEDEKAH JAKARTA," vol. 02, no. 03, pp. 429–436, 2021.
- [8] E. D. Wahyuni, "Implementasi Metode Incremental Pada Sistem Informasi Administrasi Desa Jambuwer," *J. Tekno Kompak*, vol. 15, no. 2, p. 156, 2021, doi: 10.33365/jtk.v15i2.1187.
- [9] A. Prastiyo and N. Rosmawanti, "Sistem Informasi Manajemen Tugas Akhir Berbasis Web," pp. 1095–1106.
- [10] A. A. Aziiza and A. N. Fadhilah, "Analisis Metode Identifikasi dan Verifikasi Kebutuhan Non Fungsional," vol. 3, no. 1, pp. 13–21, 2020.
- [11] J. Pendek, "Perbandingan metode cosine similarity dan jaccard similarity untuk penilaian otomatis jawaban pendek," 1978.
- [12] R. Y. Dillak, F. Laumal, and L. J. Kadja, "SISTEM DETEKSI DINI PLAGIARISME TUGAS AKHIR MAHASISWA MENGGUNAKAN ALGORITMA N-."