International Journal of Informatics Engineering and Computing (IJIMATIC)
Vol. 1, No. 1, 2024

Determining Thesis Title using Winnowing Algorithm

Rendi Saputra[1], Indah Susilawati[2]
Universitas Mercu Buana, Yogyakarta, Indonesia

Abstract
This study presents the development and application of a system utilizing the winnowing algorithm to evaluate the suitability of thesis titles based on their relevance to existing research. By generating fingerprint values from input titles and calculating similarity percentages using Jaccard's Similarity Coefficient, the system provides a quantitative measure of title uniqueness and originality. In this study, the similarity between the proposed title and existing titles was determined to be 38.4%, indicating a low level of overlap and confirming the title's appropriateness for academic research. The winnowing algorithm enhances the efficiency and accuracy of the title selection process, reducing the risk of redundancy and plagiarism while ensuring alignment with academic standards. This automated approach streamlines decision-making, enabling students to choose focused and high-quality thesis topics. The system's ability to deliver precise similarity metrics supports academic integrity and promotes originality in research. Future research could explore the integration of advanced algorithms or machine learning to further improve the system's performance and adaptability across diverse academic domains.

Keywords:
Thesis, Thesis Title, Winnowing Algorithm

1. Introduction

Thesis is one of the most important steps in college. For students, identifying a thesis title that is relevant to their interests and field of study is an important first step. However, students of Mercu Buana Yogyakarta University often have difficulty finding suitable and appropriate titles. The ineffective process in the process of submitting and receiving thesis titles so that a lot of time is wasted. In this digital era, information can be easily accessed via the Internet, which has led to a rapid increase in the number of documents available. Students should access and research various reference sources, such as scientific journals, articles and books, to gain a better understanding of the research topics they are interested in [1].

In this case, the winnowing algorithm can be a useful approach to help students identify the title of their thesis. The winnowing algorithm is a method used in text analysis to identify important parts or "fingerprints" of a document. This method has been widely used in the field of word processing and has proven to be very effective in reducing complexity and increasing the efficiency of information processing. By applying the winnowing algorithm to a collection of documents related to the research topic, students can identify and focus on the important parts of the document. This allows them to better understand the questions most relevant to their research. Thus, the Winnowing algorithm can help students save time and effort searching for documents, as well as guide them to identify more focused and quality thesis titles [2].

Corresponding Author: Rendi Saputra (191110010@student.mercubuana-yogya.ac.id)

2. Related Works

Research on the application of the winnowing algorithm, namely "Implementation of the Winnowing Algorithm in the Document Similarity Detector Application", where in this study focuses on detecting the similarity of thesis reports with other theses by giving a percentage value of similarity, after that the similarity value between articles can be measured and plagiarism can be avoided. From the results of this study, it is known that by conducting the K-fold cross validation test, the mean similarity of the dissertation of computer engineering students at Tanjungpura University by measuring the Dice Distance is 23.87% while the Chebyshev distance is 23.87%. 79 and the average time used to process documents is 110.74 seconds. The calculation results using the Chebyshev distance cannot be combined with the Dice Distance calculation results, because Dice Distance has a measurement value of 0 to 1, while the Chebyshev distance is from 0 to 1. Infinity together. If forced to combine, it will get inappropriate results [1].

Furthermore, research entitled "Implementation of the Winnowing Algorithm in Detecting Plagiarism in Student Assignments" in which this research is focused on detecting the occurrence of plagiarism in student assignments and requires a tool that can match student assignments in determining the similarity of student assignments with the winnowing algorithm. The results of the study show that the winnowing algorithm works best on tests with n-gram values (n) = 3 and window (w) = 3 [2].

The next research is entitled "Application of the Winnowing Algorithm to Detect the Similarities of Two Different Texts". In this study it focuses on assessing the similarity of texts in two different texts which requires an algorithm to assess whether the two texts have similarities. able to detect similarities between two different texts quite accurately with a percentage difference of 1% similarity [3].

The next winnowing algorithm research is entitled "Selection of the Best Parameters in the Winnowing Algorithm in Detecting the Similarity Level of Indonesian Documents". In this study, a document similarity detection algorithm was used using the fingerprint method, namely the winnowing algorithm. The winnowing algorithm has several differences in the use of parameters, such as some using k-grams and n-grams. Based on these different parameters, a performance search is carried out by comparing the use of different parameters in the string cutting process and in the process stage of the winnowing algorithm which is to determine the parameter with the best performance level. The results of the k-gram search have a high similarity value, but the higher the numerical value of k, the lower the similarity value with the average results at k = 2 of 0.5299, k = 3 of 0.1689, k = 5 of 0.0283 and k = 7 of 0.0095. The application of string cutters from n-grams to unigrams has an average similarity of 0.0683, bigram 0.003, trigram and four gram 0.000. When comparing the results of the k-gram processing speed and n-gram time, there is no significant difference and both dominate within 6 seconds [4].

The latest research is entitled "Implementation of the Winnowing Algorithm in the Automatic Assessment System for Essay Answers in Web-Based Online Examinations". This research focuses on implementing algorithms that can be used to carry out automatic assessment of online tests, especially essay questions. The winnowing algorithm is used to match data that has been identified as the answer to the system. The match rate of answers is obtained from the matching process between the system's answer key and the student's answer. Then from the results of this study it is known that the assessment of test scores by comparing system values with manual values has a fairly good level of accuracy with an average difference of 5.683% for each question [5].

This research focuses on determining thesis title which can assist students in determining thesis title by applying a similarity assessment of existing thesis titles with the winnowing algorithm. So, it is hoped that this system can assist students in determining thesis titles that are more efficient and in accordance with the desired criteria.

Theoretical basis

1. Web

   The Web is a collection of pages that provide information. [6]. Meanwhile, according to Abdullah, a website can be understood as a collection of pages that contain digital information in the form of text, images, animation, audio and video, or a combination of all of these that are sent via an internet connection so that people around the world can access and view them. Website pages are created using standard languages, especially HTML. This HTML script will be translated by a web browser so that it can be displayed as information that can be read by everyone. [7]

   According to [8], a website is an address (URL) that functions as a place to store data and information based on certain topics.

2. MySQL

   MySQL is a database software. MySQL is a relational data type, which means that MySQL stores data in the form of interconnected tables. The advantage of storing data in a database is the convenience of storing and displaying the data because it is in tabular form [9].

   According to [10]. MySQL is a DBMS application that is very widely used by web application programmers. The advantages of MySQL are that it is not paid or free, it is reliable, it is updated and there are many forums that facilitate users if they have problems. MySQL is also a DBMS that is often bundled with a web server so that the installation process is easier.

3. PHP

   PHP or short for Hypertext Preprocessor is an open-source programming language that is very suitable or specifically for web development and can be embedded in an HTML thesis. The PHP language can be said to be a description of several programming languages such as C, Java, and Perl and is easy to learn. PHP is a server-side scripting language, where data processing is carried out on the server side. Simply put, the server will translate the program script, then the results will be sent to the client who made the request [11].

   PHP programming is very suitable for development in a web environment, because PHP can be attached to HTML scripts or vice versa. PHP is specifically for dynamic web development, the meaning is that PHP is able to produce websites that can continuously change the results according to the pattern given, this depends on the request of the client browser (you can use the Opera browser, Internet Explorer, Mozilla, etc.) -other). And usually creating a dynamic web with PHP is closely related to the database as the source of the data to be displayed [12].

4. Winnowing Algorithm

   Winnowing is an algorithm used to perform the document fingerprinting process. This process is intended to be able to identify plagiarism, including small similar parts in a large number of documents. The input from the document finger-printing process is a text file. Then the output will be a set of hash values called fingerprints. This fingerprint will be used as a basis for comparison between text files that have been entered [13]. The winnowing algorithm is one of the string-matching algorithms. In its detection, the winnowing algorithm must meet the basic needs, namely [14]:

   a. Whitespace insensitivity, namely the search for similar sentences should not be affected by spaces, typeface (capital or normal), punctuation and so on.
   b. Noise suppression means avoiding finding matches with words that are too small

or less relevant, such as "the" and are not commonly used words.

c. Position independence, namely finding similarities must not depend on the position of words so that words with different positional orders can still be recognized if similarities occur.
    - Removing Irrelevant Characters. At this stage the process carried out is removing punctuation marks, spaces and symbols such as @, #,$,*,(,),!,-,_,",+,>,</ and etc
    - Formation of n-gram series: The formation of n-gram series in the winnowing algorithm is done by forming a series of characters of length n from the results of removing irrelevant characters. A good n value is neither too small nor too large. The first n-gram series starts from the 1st character to the nth character and the second series starts from the k-2nd character to the n+1st character and so on until an n-gram series of all characters is formed.
    - Hash Function Calculation for Each n-gram, The winnowing algorithm uses rolling hash to calculate the Hash value of each series of grams. The hash function with rolling hash is defined in equation 2.1.
      $H = c1 * bk –1 + c2 * bk –2 +c3 * bk –3 + ...+ ck * b$ ..... (2.1)
    - Window Formation from Hash Values, The winnowing algorithm does not use all the hash values from each series of grams formed. The hash value formed in the previous stage will be divided into windows of size w. The first window contains the first hash value to the wth hash value. The second window is formed from. The second 25 hash values up to the w+1 hash value and so on until a window of all hash values is formed.
    - Selecting the fingerprint from each window. After forming a window of all the hash values, the next step is to determine the text fingerprint value.

3. Proposed Method

The following are the stages of the methodology in this study, as shown in Figure 3.1.

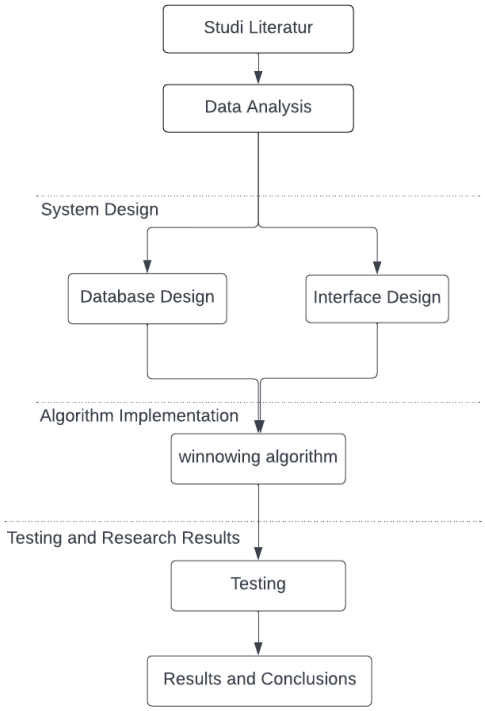

Figure 3. 1 Research Stages

Information:

1. Literature study in this study was conducted to get an overview of the research object and supporting theory in this research. Literature studies used include online journals, as well as theses related to the research needed.
2. The data analysis referred to in this study is the analysis carried out to process data into information according to what is needed in determining the title of the thesis. The first analysis is preprocessing analysis where in this analysis the process of removing unnecessary characters is carried out. Next is the analysis of the N-Grams algorithm where at this stage all the data that has been obtained will be changed from word forms to N-Grams series with the appropriate order sentence structure. Then the rolling hash analysis which analyzes changes from the form of the N-Grams series to a hash to determine fingerprinting. The final analysis is to calculate the percentage similarity of a thesis title using Jaccard's Similarity Coefficient, namely by measuring between two datasets where the result of dividing the same amount of data from the two datasets is divided by the sum of all data in the dataset. Thus, using the winnowing algorithm can determine the level of word similarity between thesis titles.
3. System design in this study is in the form of an overview that will be used in designing the main menu page of the system, namely by making database designs, and interface designs.

Database Design

Database is a system that functions as a collection of files, tables or archives that are connected and stored in various electronic media. The database used in this study uses XAMPP or PHP MySQL.

Interface Design

- System Flowcharts

The following is a flowchart of the winnowing algorithm system in determining the thesis title.

a. Input Thesis Title
   This process inputs the thesis title from students where in this process students input the title and if the title has a small percentage of plagiarism value then the thesis title will be accepted or approved as thesis title by the system and vice versa if the percentage of plagiarism value is large then the thesis title will be rejected by system.
b. Winnowing Algorithm
   In this process the system will calculate the hash value with the ASCII value for each character obtained from the n-gram series through the thesis title that has been inputted by the student, then the window formation from the hash sequence (winnowing) from the window formation is obtained from the fingerprint value obtained from the hash value smallest of the window series.
c. Jacard Similarity
   This process is carried out to calculate the percentage of similarity (similarity) of the thesis title with the formula in equation 3.1. [15]
   $D\ A,B = (|A \cap B|)/(|A \cup B|) \times 100$……..(3.1)
   Information:
   A : fingerprints value of title 1
   B : fingerprints value of title 2

- UML (Unifield Modeling Language)

Sequence diagram is a diagram that explains object interactions based on time sequence. Sequence can describe the sequence or stages that must be carried out to be able to produce something that is expected.

Usecase Diagram is a type of diagram that describes the relationship between the system and the user. Usecase can describe the type of interaction between the user and the system. Activity diagram or in Indonesian means activity diagram, is a diagram that can model the processes that occur in the system. Like a sequence of processes running a system and described vertically.

- System planning

System Design is a description, planning and sketching of various separate elements into a unified whole and functioning.

4. Algorithmic Design In research focusing on determining thesis title using the winnowing algorithm, in research focusing on determining the title of the thesis using the winnowing algorithm, as explained in equation 2.1.

4. Experimental Setup

From the results of research on determining thesis title using the winnowing algorithm, this algorithm can help determine the appropriate thesis title for students because the calculation results of this algorithm can minimize the similarity or similarity of existing thesis titles. The following is the system for determining the title of the thesis using the winnowing algorithm that runs on the system.

1. Input the thesis title you want to submit.

Figure 4.1 Entered title

From the results of the title entered, it will be compared with a title that is almost the same as the existing title first, then the two titles will be processed by removing irrelevant characters to become:

Figure 4.2 Removing Irrelevant Characters

Then, from the results of removing irrelevant characters, an n-gram value or series will be formed, where the n-gram value used in this system is n = 3. Then the title will change to the following series:

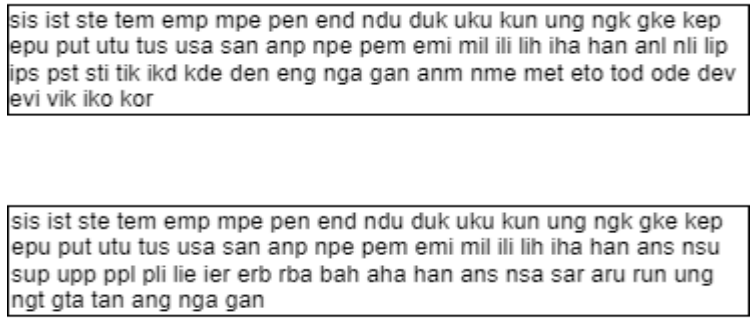

Figure 5.3 N-gram series

2. Then from the n-gram series, the hash value (rolling hash) will be calculated based on each n-gram series.
   - (sistempendukungkeputusanpemilihanlipstikdenganmetodevikor)
     sis
     $s * 11^2 + i * 11^1 + s * 11^0$
     $= (115 * 121) + (105 * 11) + 115 * 1$
     = 15.185
     Etc….
   - (sistempendukungkeputusanpemilihansupplierbahansarungtangan)
     sis
     $(s * 11^2) + (i * 11^1) + (s * 11^0)$

$= (115 * 121) + (105 * 11) + 115 * 1$
= 15.185
Etc….
From the results of these calculations, the hash value is obtained as follows:

14086 13532 13531 13494 14344 13741 13570 14955 15438 15092
13059 13525 13998 13946 13761 13055 14052 14933 13982 13321
13534 13056 13608 13329 13624 13993

14086 13532 13531 13494 14344 13741 13570 14955 15438 15092
13059 13525 13998 13946 13761 13062 14692 14892 14845 14324
13930 13573 13029 12978 13108 14559 13836 13050

Figure 4.4 Hash Value Series

3. Then we will form a window from a series of hashes (winnowing)
   From the formation of the window, we get the fingerprint of the smallest hash value of the window series.
   Table 4.1 Fingerprint of the smallest hash value

Fingerprint judul 1	Fingerprint judul 2
14086 13532 13531 13494	14086 13532 13531 13494
14344 13741 13570 14955	14344 13741 13570 14955
15438 15092 13059 13525	15438 15092 13059 13525
13998 13946 13761 13055	13998 13946 13761 13062
14052 14933 13982 13321	14692 14892 14845 14324
13534 13056 13608 13329	13930 13573 13029 12978
13624 13993	13108 14559 13836 13050

4. The last step is to calculate the results of the percentage similarity of the thesis titles that have been inputted using the Jaccard Simalarity Coefficient.

   A = (14086 13532 13531 13494 14344 13741 13570 14955 15438 15092 13059 13525 13998 13946 13761 13055 14052 14933 13982 13321 13534 13056 13608 13329 13624 13993)

   B = (14086 13532 13531 13494 14344 13741 13570 14955 15438 15092 13059 13525 13998 13946 13761 13062 14692 14892 14845 14324 13930 13573 13029 12978 13108 14559 13836 13050)

   $|A \cap B|$ = (14086 13532 13531 13494 14344 13741 13570 14955 15438 15092 13059 13525 13998 13946 13761)

$|A \cup B|$ = (14086 13532 13531 13494 14344 13741 13570 14955 15438 15092 13059 13525 13998 13946 13761 13055 14052 14933 13982 13321 13534 13056 13608 13329 13624 13993 13062 14692 14892 14845 14324 13930 13573 13029 12978 13108 14559 13836 13050)

Eyes, D(a,b) = 15/39*100= 38,4%

5. Result and Analysis

The results of the final assessment carried out by the system show how much the percentage of similarity of thesis titles is inputted by students where if the percentage of similarity of thesis titles inputted using the Jacard similarity coefficient is less than 50% then the title will be accepted by the system, conversely if the percentage of similarity is above 50% it will be automatically rejected by the system. The following are the results of the system assessment which shows the results of 38.4% of the titles that have been submitted and the title is declared accepted.

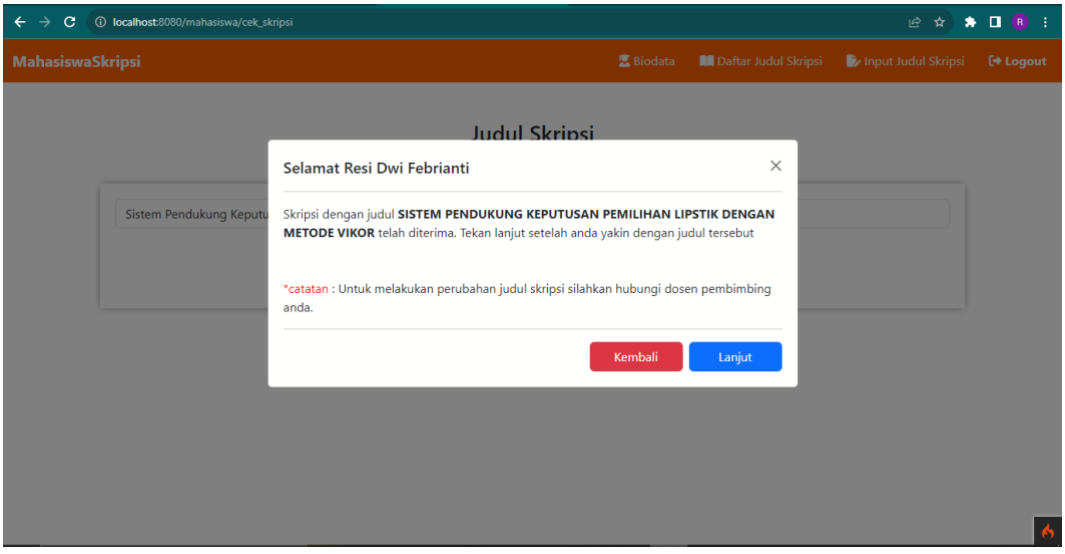

Figure 5.1 Received Title Notification

6. Conclusion

The implementation of the winnowing algorithm in this study has demonstrated its effectiveness in determining the relevance of thesis titles by analyzing their fingerprint values and calculating similarity percentages using Jaccard's Similarity Coefficient. In this case, the similarity between the entered title and existing research titles was found to be 38.4%, indicating a low level of overlap and thus confirming the suitability of the title for the student's thesis. This approach ensures that the selected title is unique and aligns with academic standards, reducing the risk of redundancy or plagiarism. By automating the evaluation process, the winnowing algorithm enhances the efficiency and accuracy of thesis title assessment, enabling a more focused and quality-driven selection process. The system's ability to provide quantitative similarity metrics supports informed decision-making, ensuring that students choose appropriate and original research topics. This

methodology not only streamlines the title selection process but also contributes to the overall integrity and quality of academic research. Future work could explore the integration of additional algorithms or machine learning techniques to further refine the system's accuracy and adaptability in diverse academic contexts.

References

[1] G. H. Mangundap, "Implementation of the Winnowing Algorithm in Document Similarity Detection Applications," *Journal of Informatics Education and Research*, pp. 147–153, 2022.

[2] I. B. Arnawa, "Implementation of the Winnowing Algorithm in Detecting Plagiarism in Student Assignments," *Journal of Information and Computers*, pp. 220–230, 2022.

[3] Jupron, "Application of Winnowing Algorithm to Detect Similarity of Two Different Texts," *Scientia Sacra: Journal of Science, Technology and Society*, pp. 246–262, 2022.

[4] W. Hidayat, "Selection of the Best Parameters in the Winnowing Algorithm in Detecting the Level of Similarity of Indonesian Language Documents," *Citec Journal*, pp. 119–132, 2020.

[5] F. E. Kurniawati, "Implementation of the Winnowing Algorithm in the Automatic Assessment System for Essay Answers in Web-Based Online Exams," *AMIK BSI Computer Engineering Journal*, pp. 169–175, 2020.

[6] E. J. Irwansyah, *Introduction to Information Technology*. Yogyakarta: Deepublish, 2014.

[7] R. Abdullah, *Web Programming For Beginners*. Jakarta: Elex Media Komputindo, 2018.

[8] Sugiyanto, "Creation of a Profile Website for the Gabus Grobogan Archipelago Development Vocational School," *Informatics and Computers*, 2013.

[9] M. A. Edy Winarno, *Web Programming Based on HTML 5, PHP, and JavaScript*. Jakarta: Elex Media Komputindo, 2014.

[10] J. K. Priyanto Hidayatullah, *WEB Programming*. Bandung: Informatics, 2017.

[11] H. F. Astria Firman, "Web-Based Online Library Information System," *E-journal of Electrical and Computer Engineering*, pp. 29–36, 2016.

[12] D. Suprianto, *Php Programming Smart Book*. Bandung: OASE Media, 2008.

[13] N. F. Ulfa and M. Mustikasari, "Creating a WEB-Based Document Similarity Level Measuring Application Using the Winnowing Algorithm," *Journal of Informatics and Computers*, pp. 61–68, 2016.

[14] Saiful, "Implementation of the Winnowing Algorithm in the Nature Encyclopedia Application," *Bulletin of Information Systems Research (BIOS)*, pp. 23–29, 2022.

[15] S. Sunardi, A. Yudhana, and I. A. Mukaromah, "Implementation of Plagiarism Detection Using the N-Gram and Jaccard Similarity Methods for the Winnowing Algorithm," *Transmission: Scientific Journal of Electrical Engineering*, vol. 20, no. 3, pp. 105–110, 2018.