



Implementation of Winnowing Algorithm and Simple Additive Weighting SAW for Publication Reference Journal Search System

Imam Fakhruddin⁽¹⁾, Indra Gita Anugrah⁽²⁾

Universitas Muhammadiyah Gresik, Indonesia

E-mail: ⁽¹⁾imam_170602@umg.ac.id; ⁽²⁾indragitaanugrah@umg.ac.id

Received: 10 May 2021; Revised: 8 September 2021; Accepted: 10 September 2021

Abstract

The Information retrieval systems are very influential in everyday life, especially in searching for reference journals. Reference journals have several different writing arrangements. In the reference journal, there is some information or words needed in making scientific journals. With this information or word retrieval system, it can help to find journals that match the similarities between input and information or words in the journal that will be used as a reference. In this study, the input process of equations and information or words will be processed using the Winnowing algorithm as an algorithm that can find the similarity of words or texts with N-gram functions, rolling hash, and Jaccard Coefficient. In general, the search only uses the same words or text without any weighting on the composition of the reference journal. To be able to find journals and their level of importance in the order of journals, a weighting method is needed. This study also uses the Simple Additive Weighting (SAW) method as a process to determine the value of the order of urgency in journals so that it can provide results in the form of rankings based on searches and urgency weights in reference journals. The results of the similarity query with documents obtained 60% precision, 77% recall, and 81% accuracy, documents, and documents had 41% precision, 83% recall, and 66% accuracy. using Winnowing Algorithm, the search system can detect the similarity of text. Determining the correct N-Gram, window, and base can increase the similarity value in the document. The document scraping process plays an important role in finding the text in each document structure. stemming can affect system performance because the steaming process helps in getting basic words (terms) in the word bag.

Keywords: Search system, Winnowing, Text similarity detection, Simple Additive Weighting

Introduction

The reuse of information nowadays has an important role, one of which is in finding reference sources used in writing research results (scientific publications). In searching for reference sources, a mechanism is needed to recognize the similarity in the text structure of a query, both in the form of keywords and documents.

Journal publications have a taxonomy or writing structure that must exist, including Title,

Abstract, Introduction, Literature Study, Methods, Discussion, Results and Conclusions, and Bibliography. A document consists of several sentences composed of several words so that we need a mechanism to detect the similarity of the query with the contents of the published document.

Several obstacles in finding references to publication documents. First, how to recognize the structure or arrangement of words so that the similarity of the sentences can be measured. The second problem with sentence similarity detec-

tion applications is that it usually only detects similarity in word structure in documents without paying attention to the weight of sentence similarities based on the level of "importance" in the taxonomy or structure of the publication document, for example, titles, abstracts, keywords, introduction, and conclusions. The taxonomic structure of text similarity of wording needs to be known in textual structure analysis (Anugrah & Sarno, 2017). So that it can make it easier to get information retrieval.

This study focuses on the application of the method of measuring text similarity using the Winnowing Algorithm and considering the weight of importance in the taxonomy of publication documents using the Simple Additive Weighting (SAW) method is able to produce the expected publication journal reference search system.

The winnowing algorithm is a method to find the similarity of text structures or words in the text structure by forming grams in each window of the hash value. In the next step, the similarity will be calculated using the Jaccard Coefficient Similarity so that it provides a percentage value for the similarity of text or documents.

The Simple Additive Weighting method is an attribute decision-making method used to obtain the best alternative by determining criteria and creating a normalization matrix and calculating the total alternative value that will be used as a ranking value.

The result of applying the proposed method is the ranking of scientific publication documents based on the similarity between queries or documents with scientific journal documents in the publication repository.

Previous research (Wibowo & Hastuti, 2016) concluded that the Winnowing Algorithm is one of the algorithms capable of detecting the similarity of text structures in a fairly fast time. The weakness of this algorithm is that it cannot provide guarantees and evidence against similar texts.

Several studies were previously conducted (Sunardi et al., 2018) The results of the above research can be concluded that n-gram greatly affects the results of similarity.

The use of the correct n-gram greatly affects the percentage of similarity. To solve the

problem of important value in the structure or taxonomy of publications, it is possible to use the Simple Additive Weighting (SAW) method such as (Mukroji et al., 2019).

SAW is one of the attribute decision-making methods, this method is usually used to get the best alternative by determining the criteria and calculating the total value of the alternative that will be used as a decision-making method. Ranking values (Mukroji et al., 2019).

The Researcher Gives the conclusion that the SAW algorithm is able to provide better results than the alternative WP ranking that appears low. SAW than WP, because the algorithm is proven to be able to minimize the same preference value, so that alternative ranking can be done well (Berlilana et al., 2018).

From several previous studies, it can be assumed or hypothesized that the Winnowing Algorithm and Simple Additive Weighting can be used for document searches by measuring the similarity of text structures according to the taxonomic importance weight of the publication journals which will later be ranked according to the entered keywords.

Materials and Methods

This Search for publication journals is a form of information retrieval system using a winnowing algorithm that can detect the similarity of contents in reference text in documents compared to similar proportions of several words or sentences. The Simple Additive Weighting method is ratings on alternative weights. The results of weighting can produce the best order (Berlilana et al., 2018).

This study will use a combined method of measuring the similarity of text structures using the Winnowing Algorithm and the calculation of the ranking of publication journal reference documents using the Simple Additive Weighting (SAW) method. At the initial stage of uploading documents, text extraction will be carried out on the document based on taxonomy or journal publication structure. The journal publication structure used is the title, abstract, keywords, introduction and results. Furthermore, pre-text processing will be carried out which aims to obtain the basic words that compose the document.

Then the measurement of the similarity of the

text structure is carried out and the final step is to calculate the ranking of the publication journal reference document. An overview of the flow of the research method is shown in Figure 1. In this study using the N-Gram setting with $N = 8$, Prime number (Basis) $B = 2$, Window $W = 7$. A high N-Gram value will affect the completion of letters in the text so that the text the less detailed but the results will be close to the same similarity, on the other hand, if the N-Gram value is small, the letters taken will be less and more detailed so that it affects the percentage value of the search. For the window value, if the window value is high, fewer results will be found and if the window is small, more results will be found so that it can affect the similarity level in the percentage of document searches. Figure 1 shows the process flow

The system starts with uploading documents, then continues with the document scraping process. In the scraping process, the text is extracted from the document in pdf format.

The next step is text pre-processing in-text pre-processing, filtering the text using case-folding to normalize the text, then stopping word deletion to remove irrelevant words and stemming literature to convert sentences into basic words then continue the process of searching. the similarity of text to Wnnowing Algorithm. contained in the formation of N-Gram, N-Gram is the stage of breaking each character from the N value in the text, then making the ASCII value for each character from N-Gram after that the rolling hash calculation process is carried out to get the value of each gram and continue the process Window formation which is determined from the value of W as the number of windows per row formed by rolling hash results, followed by the fingerprint process, in this process the smallest rolling hash value of each row of windows will be found so that the fingerprint value is found in every document. In the search section, a process similar to uploading a document is carried out but without scraping. Search does not require scraping as text can be retrieved immediately.

The next step is to calculate the value generated from the document fingerprint and the fingerprint from the query to get the total text similarity value based on the taxonomy of the

document being searched using the Jaccard Coefficient. Because there is a total similarity value that is taken based on the taxonomy of the similarity results, the ranking of values will be carried out using the Simple Additive Weighting (SAW) method to find the largest document value based on weight or taxonomic structure. You can also search for similarities in the selected documents based on the query results found. So that the results of document similarities will appear.

Wnnowing Algorithm

Wnnowing algorithm is a method of finding text similarities. This method is an extension of the Rabin-Karp Fingerprint method with window characteristics on the process and good results. The Wnnowing algorithm looks for word equivalence (document fingerprinting) to find similar text. The wnnowing algorithm is quite reliable because it can look for text similarities in words (Sibarani et al., 2019). Selection of the smallest fingerprint value results in each window that will be calculated by the equation with Jaccard Coefficient.

Following are the steps for the Wnnowing Algorithm:

Word Combining

At this point, all the words will be combined without spaces.

N-gram process

N-gram is a series on wnnowing by splitting the length of the character = n. the initial sequence starts from the 1st to the nth character, then the next character starts from the 2nd to the nth + 1 and so on, after that the chained characters will be hashed by rolling the hash process.

Rolling Hash Process

Make ASCII per character based on a series of grams and calculate the value of ASCII characters, the formula for the rolling hash is as Equation (1).

$$c_1 * b^{k-1} + c_2 * b^{k-2} * \dots + c_{k-1} * b + c_k \quad (1)$$

Information :

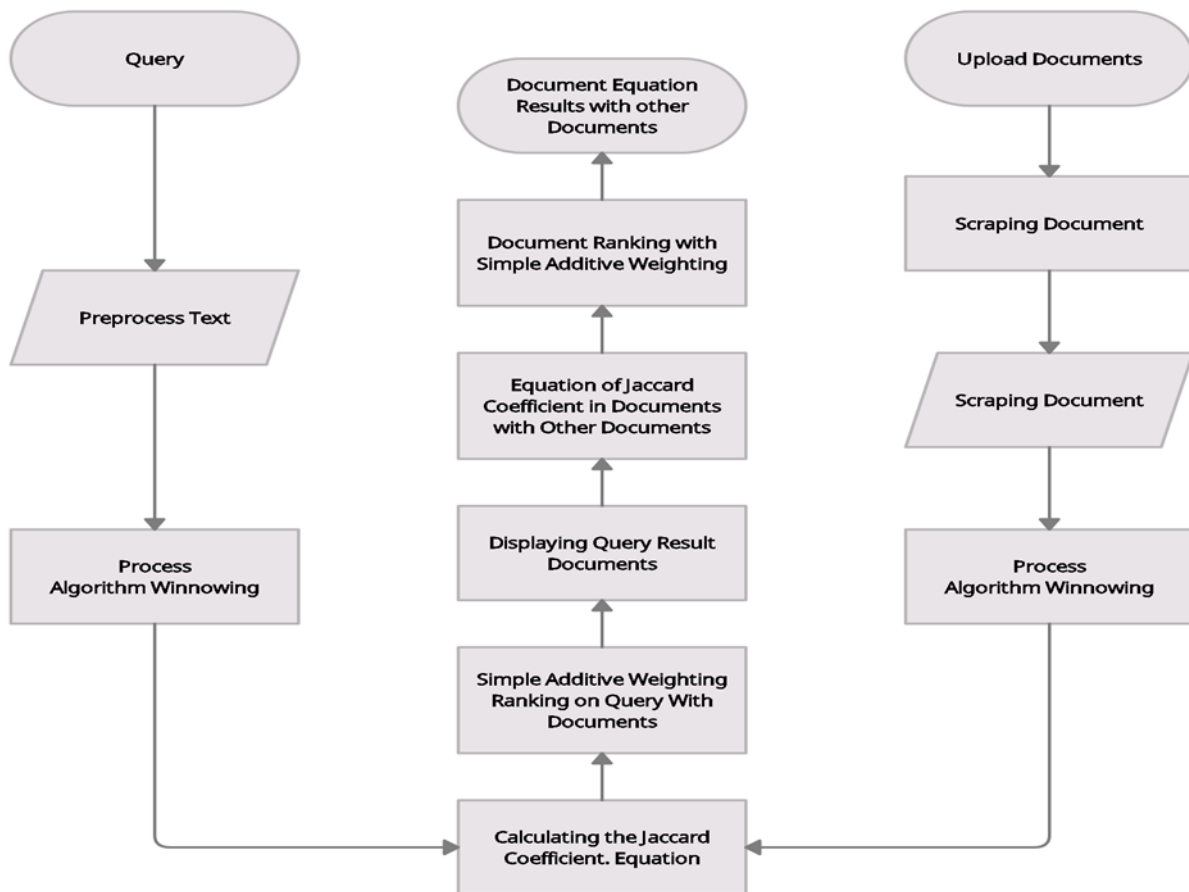


Figure 1. Research Diagram

c = Ascii character value, b = prime number,
and k = gram value.

Process Window

The resulting window formation from each gram becomes a hash value with was the window size, the 1st window contains the hash value from the first gram continued to the nth window, and for the second window 2 to the n + 1 window and so on.

Fingerprint Process

After the window is complete, the next step is to take each smallest window value from each row into a fingerprint, if there are two or more smallest values of the same fingerprint, only one will be taken that is on the right, from the results of the window will look for the smallest value for each window.

The Jaccard Coefficient Equation

The results of the winnowing algorithm will form a fingerprint value, this fingerprint value will be used as the level of text similarity in the Jaccard coefficient as Equation (2):

$$\text{Similarity} = \frac{\text{jumlah } h \text{ fingerprint sama}}{\text{total seluruh } h \text{ fingerprint}} \times 100\%$$

Or

$$\text{Similarity}(di, dj) = \frac{|w(di) \cap w(dj)|}{|w(di) \cup w(dj)|} \times 100\% \quad (2)$$

information :

fingerprint value on text,

dj fingerprint values in the text $w(di) \cap w(dj)$

the number of fingerprint values is the same between the with text and

jth text and $w(di) \cup w(dj)$ is the total value fingerprint ith text and jth text.

Simple Additive Weighting

Simple Additive Weighting (SAW) is a decision-making method used to determine the best alternative. The weighting criteria that exist in Simple Additive Weighting (SAW) will affect the value at the time of the ranking process, making matrices, normalizing the matrix, and

calculating the total value on the criteria whose value has been determined. The Simple Additive Weighting (SAW) method is combined with the winnowing algorithm because this study uses several values from the taxonomy of titles, abstracts, keywords, introduction and results in documents so that it requires an alternative to being able to get documentary values which will be ranked according to the highest value (Alamsyah, 2017) as Equation (3).

$$R_{ij} = \frac{\frac{x_{ij}}{\max x_{ij}}}{\frac{\min x_{ij}}{x_{ij}}} \quad (3)$$

R_{ij} : Normalized performance rating value

x_{ij} : The attribute value that each criterion has

$\frac{x_{ij}}{\max x_{ij}}$: The largest value of each criterion

$\frac{\min x_{ij}}{x_{ij}}$: The smallest value of each criterion

Benefit: If greatest value is best. Cost: If smallest value is best.

Result and Discussion

Scraping Documents

Document scraping is a way to retrieve information by retrieving the contents of the text in the file, by providing initial and final limits that have been determined according to the structure of scientific publication journals, namely "Title", "Abstract", "Keywords", "Introduction" and "Results". Here's the scraping source code, shown in Figure 2:

```
$source_pdf="jurnal/$name_file";
$output_folder="MyFolder/jurnal";
if (!file_exists($output_folder)) {
    mkdir($output_folder, 0777, true); }
$a= passthru("pdftohtml $source_pdf $output_folder/$name_file",$b);
$myfile = fopen("MyFolder/jurnal/$name_file$.html" ,"r") or error($name_file);
$text= strtolower(fread($myfile,filesize("MyFolder/jurnal/$name_file$.html")));
```

Figure 2. Document Scraping Script

Preprocess Text

This process performs text filtering from scraping results. Text will be normalized to be uniform, for example, lowercase letters with lowercase, removing punctuation or symbols, meaningless words will be deleted then the stemming process is carried out using stemmer literature so that a good root word is obtained. The following is the pre-text stage:

Case Folding

At this stage, the text normalization process is carried out, converting all text to lowercase, deleting punctuation marks, deleting symbols, and deleting numbers such as symbols @ # \$! * & ^% <? / : , 123456789 ". Case folding is done to produce more consistent text.

Stopword Removal

After the case folding process is carried out at this stage, words that have no meaning, do not provide important information or words that cannot stand alone, for example, 'yang', 'di', 'ke', 'because', etc. (Anwar et al., 2019)

Stemming

The sentence produced after the case folding and stopwords removal process is a sentence that still contains affix words. Words that have an affix cause a root word to have a double meaning so the stemming process must be carried out. The stemming process is a mechanism for generating basic words by eliminating prefixes, insertions, and suffixes. The stemming process must consider the grammatical structure, for example, words from Indonesian or English. In this study, the document used is a published journal document in Indonesian so that the stemming method uses a stemmer for Indonesian. Literary stemmer is a stemmer used for Indonesian words. The literary library has a dictionary obtained from kateglo.com. kateglo.com is an Indonesian dictionary website (Anwar et al., 2019). Figure 3 is an example from a literary stemmer:

```
"#Pertandingan Final Pada Liga Eropa@ Antara
Tim A Melawan Tim B Yang Berlangsung
Dengan Hasil Tim A Sebagai Juara!!"
```

Figure 3. Not Preprocessed Text

tanding final liga eropa antara tim a lawan tim b
langsung hasil tim a bagai juara

Figure 4. Preprocessed Text

tandingfinalligaeropaantaratimalawantimblang-
sunghasiltimabagaijuara

Figure 5. Remove Spaces

Winnowing Algorithm

After preprocessing the text of the sentence as shown in Figure 3 and the results of the sentence are shown in Figure 4. The sentence that has been pre-processed will continue the process using the Winnowing Algorithm. The application of the Winnowing Algorithm in this example will be carried out using the N-gram value $N = 3$, Basis (prime number) $= 2$, and Window $W = 7$ to look for the similarity value of the text structure. The thing that must be done before looking for similarities in text structures using the Winnowing Algorithm is to remove sentence spacing so that it will produce the text in Figure 5.

Solving N-Gram sentences

From the results of combining the text, the text will be split with N-Gram $N = 3$. Shown in Figure 6.

tan and ndi din ing ngf gfi fin ina nal all lli lig iga
gae aer ero rop opa paa aan ant nta tar ara rat ati
tim ima mal ala law awa wan ant nti tim imb mbl
bla lan ang ngs gsu sun ung ngh has asi sil ilt lti
tim ima mab aba bag aga gai aij iju jua uar ara

Figure 6. N-Gram results

Rolling Hash Process

After the sentence is broken down in the N-Gram process, then the characters on the N-Gram are converted into ASCII per character based on the Gram series and then the ASCII character values will be calculated using the formula from Rolling Hash with Equation (1). The example in Figure 7 is the Rolling Hash calculation from Gram that has been formed previously from Figure 6. The example in Figure 7 grams taken is "tan".

$$= t * 2^{(\text{string length} - 1)} + a * 2^{(\text{string length} - 2)} + n'$$

$$= 116 * 4 + 97 * 2 + 110'$$

$$= 464 + 194 + 110'$$

$$= \mathbf{768} \text{ (and more).}$$

Figure 7. Rolling Hash calculation from Gram

Process Window

Process Window is a segment formation mechanism that contains a collection of Rolling Hash processes. In this example, a Window value of 7 is used, so that the segment shown in Figure 8 will be 7 rows.

768 708 745 720 743 748 721
708 745 720 743 748 721 728
745 720 743 748 721 728 737
720 743 748 721 728 737 742
(..... and more)

Figure 8. Segment Representation

Fingerprint

After segmenting using Window, it will be continued by finding the smallest value from each window in the Fingerprint process. Figure 9 and Figure 10 are the processes of finding the smallest value from the Fingerprint Query with the Fingerprint of each Document. The query in this example after pre-processing the text is "The final match of the European league between team a against team b directly results from the team alike champions" and the document in this example is "team a won the European league after defeating team b in the final match last night".

Fingerprint Query =
[693 709 707 704 717 703 701 689 708 720]

Figure 9 Fingerprint from Query

Fingerprint Document =
[708 720 712 707 704 692 713 701 723 705 711
681 689 691]

Figure 10. Fingerprint of the Document

Fingerprint Query = [693 709 **707704** 717
703 **701689708720**]

Fingerprint Document = [**708720** 712 **707 704**
692 713 **701** 723 705 711 681 **689** 691]

Figure 11. The similarity of Fingerprint Value between Query and Document

Similarity Jaccard Coefficient

The last stage of the Winoing Algorithm is to calculate the similarity between the Fingerprint Query and the Fingerprint of each Document using the Jaccard Coefficient.

At this stage, the similarity value between the Query and the Document from Figure 11 will be calculated. Then Union Fingerprint Query with Documents: 708 720 707 704 701 689 = 6 is the same number between the Fingerprint Query and the Document Fingerprint. The total of Fingerprints that do not include Union is: 12 and added from the results that are included in Union are: 6. The total is 18. So that the result of the Measurement of the Similarity between Fingerprint Query and Documents: $(6/18) \times 100 = 33.33\%$ Equation (2).

Simple Additive Weighting (SAW)

From the measurement results of the similarity of the text structure using the Winoing Algorithm which is carried out on each taxonomy journal publication. In this study, the taxonomies used were title, abstract, introductory keywords, and results. From each of these taxonomies, the value will be calculated

Table 1. Weights of Publication Journal Taxonomy Preference

Title	Abstract	Keywords	Intro	Results
0,1	0,275	0,25	0,175	0,2

according to the preference weight in Table 1 using Simple Additive Weighting (SAW) for ranking documents. The following formula is designated in Equation (3).

Formation of Matrix X

The Simple Additive Weighting (SAW) method begins with the creation of an M x N matrix of the suitability of each alternative on the criteria (Supriyatna & Ekaputra, 2017). where M is the number of scientific journal publication document data and N is the taxonomic number of scientific journal publication documents. Table 2 is an example of the Matrix generated from the results of the Jaccard Coefficient equation.

SAW Ratings

After the R Matrix (Normalized Matrix) is produced, the publication process will be carried out by using Equation (3) so that a document ranking is produced from the largest to the smallest value. The greater the value of the calculation result, the higher the ranking compared to the smaller value. The following Table 3 is the result of the Simple Additive Weighting (SAW) ranking from the results of the R Matrix calculation and taxonomic weights in Table 1.

Table 2. X (M x N) Matrix of Publication Document Data

Doc	Taxonomy of Publication Journal Structure				
	Title	Abstract	Keywords	Introduction	Results
Doc 1	29,31	64,51	15	78,70	65,71
Doc 2	32,72	54,41	14,70	82,72	71,81
Doc 3	36,36	61,76	14,39	82,35	74,07
.....					
Doc 68	31,70	44,11	9,09	76,31	83,92
Doc 100	34,14	61,29	10	80,35	80,50

Table 3. Document Ranking Results Using SAW

Document	SAW
Document 3	0,9254
Document 1	0,8943
Document 2	0,8795
Document 4	0,8073

To find out the results of document ranking using SAW, it is necessary to carry out manual testing by experts. Manual testing is done by matching the text of each sentence in the document from the rating results using SAW by matching the text of each sentence by the expert manually. Figure 14 is the test result.

System Implementation

From the Implementation of the Wining Algorithm and Simple Additive Weighting (SAW) Application System in the Search for Publication Journal Reference and the following system interface:

Upload documents

The document upload page is used for users who want to save results from previous journal searches from the internet and carry out the scraping, pre-processing, and winnowing algorithm processes. Shown in Figure 12.

Display search and search results

The search page is on the same page as the upload in Figure 13 which is the query input for the search. The search results contain documents from the document upload repository by performing a winnowing algorithm process as a detection of text similarities and ranking using the simple additive weighting method so that results are obtained on document contents that are most similar to the input in the query. Shown in Figure 14.

Document view by document

On this page are the results of the search for document similarity to the document in Figure 14. Figure 15 is the results from document to document.

Testing

To find out the results of the recommendation system for finding a publication journal reference document, the proposed method is a combination of the Winnowing Algorithm and Simple Additive Weighting (SAW). The test will be carried out with 4 different queries for the document. The standard value of the 40 existing documents will be taken using the threshold formula (Francq, 2014). So that every query that is entered has a different standard value. The recommended search results will be calculated and the system output results are matched with the user (Human). Precision, Recall, and Accuracy values will be calculated using the Confusion Matrix (Falsini et al., 2012).

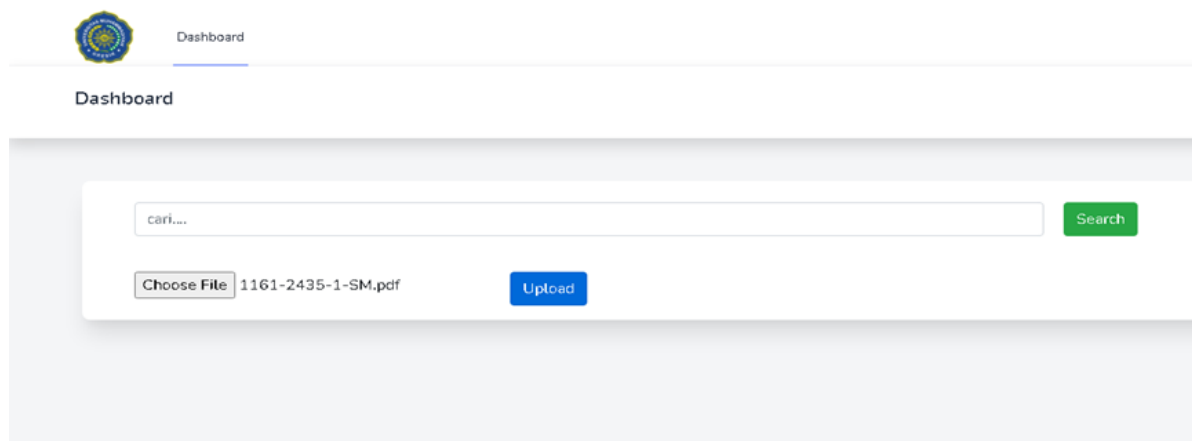


Figure 12. Upload the document

PENGARUH PERSEPSI SISWA TENTANG METODE MENGAJAR GURU

Search

Choose File No file chosen

Upload

Figure 13. Search

Query : PENGARUH PERSEPSI SISWA TENTANG METODE MENGAJAR GURU

Show 10 entries

No	Document	Similarity Result	Action
22	[jurnal pendidikan akuntansi indonesia vol xi no tahun artin nosa sefrian sari amp abdullah taman halaman pengaruh motivasi berprestasi dan persepsi siswa tenta	0.8975	View Document Relevansi
2	[tugas dan fungsi kepolisian dalam perannya sebagai penegak hukum menurut undang undang nomor tahun tentang kepolisian muhammad arif universitas islam kalimantan jt	0.8771	View Document Relevansi
36	[jurnal reka karsa jurusan teknik arsitektur itenas no vol jurnal online institut teknologi nasional oktober kajian pola	0.821	View Document Relevansi
28	[jurnal pendidikan akuntansi indonesia vol x no tahun arga lacopa arisana amp ismani halaman pengaruh kedisiplinan siswa dan persepsi siswa tentang kualitas mengaja	0.7861	View Document Relevansi
39	[see discussions stats and author profiles for this publication at https www researchgate net publication mengembangkan soft skill siswa melalui proses pembelajaran articl	0.7845	View Document Relevansi
24	[issn e issn x economica journal of economic and economic education vol no pengaruh persepsi siswa tentang gaya mengajar guru dan minat belajar	0.7763	View Document Relevansi

Figure 14. Search results

No	Document	Similarity Result	Action
19	jurnal pendidikan akuntansi indonesia vol xi no tahun artin nosa sefrian sari amp abdullah taman halaman pengaruh motivasi berprestasi dan persepsi siswa tentan..... === perancangan aplikasi sistem pakar penyakit kulit pada anak dengan metode expert system development life cycle annisa nurul fadhilah dini destiani dhami johar dhamiri jurn.....	0.8469	View Document
28	jurnal pendidikan akuntansi indonesia vol xi no tahun artin nosa sefrian sari amp abdullah taman halaman pengaruh motivasi berprestasi dan persepsi siswa tentan..... === journal of mechanical engineering education vol no desember model pembelajaran teams games tournament untuk meningkatkan hasil belajar siswa smk wisnu d yudian.....	0.8051	View Document
13	jurnal pendidikan akuntansi indonesia vol xi no tahun artin nosa sefrian sari amp abdullah taman halaman pengaruh motivasi berprestasi dan persepsi siswa tentan..... === http jurnal fk unand ac id artikel penelitian gambaran faktor risiko dan tipe stroke pada pasien rawat inap di bagian penyakit dalam rsud kabupaten so.....	0.7976	View Document
11	jurnal pendidikan akuntansi indonesia vol xi no tahun artin nosa sefrian sari amp abdullah taman halaman pengaruh motivasi berprestasi dan persepsi siswa tentan..... === jurnal riset sistem informasi dan teknik informatika jurasik volume no juli issn eissn http tunasbangsa ac id ejurnal index php jurasik	0.7828	View Document
24	jurnal pendidikan akuntansi indonesia vol xi no tahun artin nosa sefrian sari amp abdullah taman halaman pengaruh motivasi berprestasi dan persepsi siswa tentan..... === jurnal maju jurnal pendidikan matematika vol no maret september hlm meningkatkan prestasi belajar matematika melalui pembelajaran kooperatif tipe jigsaw pada siswa sm.....	0.7758	View Document

Figure 15. The results of document similarity to the document

Table 6. Test Results With Query

No	Query	With Query		
		<i>Precision</i>	<i>Recall</i>	<i>Accuracy</i>
1	Hubungan Antara Gaya Mengajar Guru Dengan Minat Belajar Pada Siswa	75%	90%	90%
2	Penataan Bentuk Bangunan Tradisional Pada Rumah Adat	64%	90%	85%
3	Tindak Pidana Sumber Hukum Penegak Hukum Hak Asasi Manusia	50%	70%	75%
4	Risiko Terhadap Kejadian Penyakit Pada Diagnosa	54%	60%	77%
	Average	60%	77%	81%

Table 7. Table Testing With Documents by Documents

No	Document by Document	With Documents		
		<i>precision</i>	<i>recall</i>	<i>accuracy</i>
1	Hubungan Antara Teaching Style (Gaya Mengajar Guru) Dengan Minat Belajar Matematika Pada Siswa	45%	100%	70%
2	Struktur Konstruksi Arsitektur Tradisional Bangunan Tradisional Keda Suku Ende Lio Di Permukiman Adat	41%	77%	67%
3	Asas Legalitas Dalam Hukum Pidana Nasional Dan Hukum Pidana Internasional	47%	88%	72%
4	Sistem Pakar Untuk Mendiagnosa Penyakit Saluran pencernaan Menggunakan Metode Dempster Shafer	31%	66%	57%
	Average	41%	83%	66%

From table 6, the test results obtained the percentage value of similarity and ranking on taxonomic weights. The average value of precision is 60%, recall of 77%, and accuracy of 81% of documents that have TP (True Positive) and FP (False Positive) values. From table 7, the test results of documents with documents obtained an average value of 41% precision, 83% recall, and 66% accuracy of documents that have TP (True Positive) and FP (False Positive) values.

The recall value on the similarity of the document to the document has a higher value than recall on the similarity of the query to the document. Because recall in document and document compares the entire contents of each structure of each document. Meanwhile, the similarity of the query to the document only compares the text content of the query with the overall contents of each structure of each publication journal document. This is proportional to the precision performance because the more contents of the text are compared, the greater the similarity value is possible.

Conclusion

From the test results, the following conclusions were obtained: (1) In testing the similarity system using a query with documents, the results are 60% precision, 77% recall, and 81% accuracy, while the similarity of documents with documents has 41% precision, 83% recall, and 66% accuracy. (2) By using the Winnowing Algorithm the search system can detect text similarities. In research conducted using N-Gram 8, Window 7, and basis 2 and the application of the Simple Additive Weighting method helps document ranking based on the structure of published journal documents. (3) Determining the correct N-Gram, window, and basis in the Winnowing algorithm can increase the similarity value (similarity) in the document. (4) The process of scraping documents plays an important role in the retrieval of text in each journal publication document structure which is used as data in the document corpus. (5) The use of stemming algorithms can affect system performance because the stemming process helps in getting the basic words (terms) in the bag of words.

Suggestion

In this study using Indonesian language journals, it is hoped that further development can process words or texts with the English stemming method so that they can filter English words and can find better N-gram values and rolling hashes for settings on the winnowing algorithm used to increase precision in document search.

Rereferences

- Anugrah, I. G., & Sarno, R. (2017). Business Process model similarity analysis using hybrid PLSA and WDAG methods. *Proceedings of 2016 International Conference on Information and Communication Technology and Systems, ICTS 2016*, 231–236. <https://doi.org/10.1109/ICTS.2016.7910304>
- Wibowo, R. K., & Hastuti, K. (2016). Penerapan Algoritma Winnowing Untuk Mendeteksi Kemiripan Text pada Tugas Akhir Mahasiswa. *Techno.Com*, 15(4), 303–311.
- Sunardi, S., Yudhana, A., & Mukaromah, I. A. (2018). Implementasi Deteksi Plagiarisme Menggunakan Metode N-Gram Dan Jaccard Similarity Terhadap Algoritma Winnowing. *Transmisi*, 20(3), 105. <https://doi.org/10.14710/transmisi.20.3.105-110>.
- Syahputra, A. R. (2017). Implementasi Algoritma Winnowing Untuk Deteksi Kemiripan Judul Skripsi Studi Kasus: STMIK Budi Darma. *Pelita Informatika Budi Darma*, 12(1), 1–9.
- Alamsyah, N. (2017). Perbandingan Algoritma Winnowing Dengan Algoritma Rabin Karp Untuk Mendeteksi Plagiarisme Pada Kemiripan Text Judul Skripsi. *Technologia: Jurnal Ilmiah*, 8(3), 124. <https://doi.org/10.31602/tji.v8i3.1116>.
- Mukroji, M., Munadi, R., & Syahrial, S. (2019). Implementasi Metode Simple Additive Weighting Pada Penentuan Peringkat Kerentanan Website (Studi Kasus Website Universitas Negeri Aceh). *Informatika Mulawarman : Jurnal Ilmiah Ilmu Komputer*, 14(1), 7. <https://doi.org/10.30872/jim.v14i1.1431>.
- Berlilana, B., Prayoga, F. D., & Utomo, F. S. (2018). Implementasi Simple Additive Weighting dan Weighted Product pada Sistem Pendukung Keputusan untuk Rekomendasi Penerima Beras Sejahtera. *Jurnal Teknologi Informasi Dan Ilmu Komputer*, 5(4), 419. <https://doi.org/10.25126/jtiik.201854768>.
- Kurniawan, D., Aristoteles, & Amirudin, A. (2015). Pengembangan Aplikasi Sistem Pembelajaran Klasifikasi (Taksonomi) dan Tata Nama Ilmiah (Binomial Nomenklatur) pada Kingdom Plantae (Tumbuhan) Berbasis Android. *Jurnal Komputasi*, 3(2), 120–128.
- A. Yani, D. D., Pratiwi, H. S., & Muhandi, H. (2019). Implementasi Web Scraping untuk Pengambilan Data pada Situs Marketplace. *Jurnal Sistem Dan Teknologi Informasi (JUSTIN)*, 7(4), 257. <https://doi.org/10.26418/justin.v7i4.30930>
- Anwar, M. S., Much, I., Subroto, I., & Mulyono, S. (2019). Sistem Pencarian E-Journal Menggunakan Metode Stopword Removal Dan Stemming, 58–70.
- Sibarani, L., Magdalena, M., & Dharma, A. (2019). Analisa Perbandingan Sistem Pendeksian Kemiripan Judul Skripsi Menggunakan Algoritma Winnowing Dan Algoritma Rabin Karp. *REMIK (Riset Dan E-Jurnal Manajemen Informatika Komputer)*, 4(1), 69. <https://doi.org/10.33395/remik.v4i1.10174>
- Ro'is, M. A. (2018). Implementasi Metode Jaccard Similarity Pada Aplikasi Pencarian Lirik Lagu. *Artikel Skripsi Universitas Nusantara PGRI Kediri*, 6.
- Setiadi, A., Yunita, Y., & Ningsih, A. R. (2018). Penerapan Metode Simple Additive Weighting(SAW) Untuk Pemilihan Siswa Terbaik. *Jurnal Sisfokom (Sistem Informasi Dan Komputer)*, 7(2), 104. <https://doi.org/10.32736/sisfokom.v7i2.572>
- Supriyatna, A., & Ekaputra, D. (2017). Metode Fuzzy Simple Additive Weighting (Saw) Dalam Pemilihan Ketua Osis. *Jurnal PE-TIR*, 10(1), 71–76.
- Budi, S. (2012). Pembangunan sistem pendukung keputusan usaha, 8–13.
- Diego Falsini, Federico Fondi & Massimiliano M. Schiraldi (2012) A logistics provider

evaluation and selection methodology
based on AHP, DEA and linear program-
ming integration, International Journal of
Production Research, 50:17, 4822-4829,
DOI: 10.1080/00207543.2012.657969.

Francq, Pascal. (2014). Re: Determination of
threshold for cosine similarity score?. Re-
trieved from: [https://
www.researchgate.net/post/Determination
-of-threshold-for-cosine-similarity-
score/52ea1b58d5a3f2c3768b45fd/
citation/download](https://www.researchgate.net/post/Determination-of-threshold-for-cosine-similarity-score/52ea1b58d5a3f2c3768b45fd/citation/download).