Journal of Engineering and Applied Sciences 13 (Special Issue 10): 8119-8125, 2018
ISSN: 1816-949X

Implementation of Winnowing Algorithm Based K-Gram to Identify Plagiarism on File Text-Based Document

Yanuar Nurdiansyah and Fiqih Nur Muharrom
Sistem Informasi Program Studi Sistem Informasi Universitas Jember (UNEJ),
Jln. Kalimantan 37, 68121 Jember, Indonesia

Abstract: Plagiarism occurs when the students have tasks and pursued by the deadline. Plagiarism is considered as the fastest way to accomplish the tasks. This reason makes the research tried to build a plagiarism detection system with winnowing algorithm as document similarity search algorithm. The documents that being tested are Indonesian journals with extension .doc, .docx and txt. Similarity calculation process through two stages, the first is the process of making a document fingerprint using winnowing algorithm and the sec is using Jaccard coefficient similarity. In order to develop this system, the research used iterative waterfall model approach. The main objective of this project is to determine the level of plagiarism. It is expected to prevent plagiarism either intentionally or unintentionally before our journal published by displaying the percentage of similarity in the journals that we make.

Key words: Plagiarism, winnowing algorithm, jaccard's coefficient, fastest way, fingerprint, determine

INTRODUCTION

The rapid technology improvement does not only give positive effect for daily life but also, gives negative influences that cannot be avoided. In this era, information is easy to get and influence people. One of the negative impacts is the great number of plagiarism. Plagiarism often occurs when the students have tasks and pursued by the deadline. The students consider that conducting plagiarism is the fastest way to accomplish the assignment.

Students regard that doing plagiarism is the instant way to accomplish their assignments (Anonymous, 2010). Those factors influence the result of the evaluation. This issue also, make an obstacle for teachers in detecting plagiarism regarding the numerous task file and the scientific research (IUE, 2012). One of the functions of technology is that it can be used as an alternative to identify plagiarism regarding that has been so, many scientific research published in electronic. Thus, the identifying process becomes much easier.

The technology is used which means the researcher uses the computer application to do the calculation of text-document similarity. The algorithm selection on text-document similarity search is very affecting. A mistake in algorithm selection may decrease the accuracy of document similarity calculation. Winnowing algorithm is a method of word similarity search in a document by comparing the fingerprint on the document (Cornic, 2008), the algorithm input is the text document which is processed and result in an output in the form of hash value. That hash value is then called as the fingerprint. This is the fingerprint which is used to compare the similarity of each document.

There are some basic needs that are used by algorithm in detecting the document similarity. The basic needs which have to be accomplished by detection algorithm are (Scheleimer *et al*., 2003; Oetsch *et al*., 2010).

Whitespace insensitivity, it is a word search which is not influenced by space, punctuation, type of letter (capital or normal) and etc.,

Noise suppression, the function is to avoid the finding of short word and not the common word such as 'the'

Position independence, it is a similarity finding that does not have to depend on the word position, so, words with different order still can be recognized if there is a similarity. Generally, the research principles of the algorithm in the document-similarity detection are; (in a few steps):

- Both the target text and the original text are assumed as string s with the length t
- Conducting the cleaning of punctuation, space and etc. which refer to the basic needs of detection algorithm

Corresponding Author: Yanuar Nurdiansyah, Sistem Informasi Program Studi Sistem Informasi Universitas Jember (UNEJ), Jln. Kalimantan 37, 68121 Jember, Indonrsia

- Dividing the document to be k-gram that is as parameter selected by the user. k-gram is the sub-string that is alongside with the length of character k
- Finding hash value from each k-gram
- Selecting some hash result to be the fingerprint document

The difference of algorithm and another algorithm of similarity detector is in the selection process of its fingerprint. The result of hash value calculation is divided into window w in which the smallest value will be taken from each window to be the document fingerprint.

MATERIALS AND METHODS

This research used the designing method and system development method. To do that method, we use software of life cycle development by adopting the iterative waterfall model.

System requirements analysis: The analysis system is needed at the early stage in designing and developing the system to determine the needs of plagiarism detection system that will build. On this stage, the researcher conducted literature study that learnt the retrieval information system and fingerprint document method using algorithm to identify the similarity of text document through all sorts of media, those are the internet, journals and books refer to plagiarism and text processing to get the view of algorithm that is going to be used on the system as the algorithm to detect the document similarity and what data needed to build that system.

algorithm as algorithm of document similarity search: The selected result of algorithm as the algorithm of document similarity search was based on the accuracy stage and the time speed of the process by referring to the previous research that is written on the bibliography. The stage of algorithm process that is implemented on the system is described in Fig. 1.

Data needed by system: The data used to build the system of plagiarism detection is journald data. The journal data of the system will be renewed by both administration and student as a user non-administrator. However, the student can only upload a journal which has been approved by the administrator. If the journal is already approved, it can be saved in the database and used as the reference by other students for next verification in order to avoid the damage of the journal data in the database. Database for the journal is divided into two table.

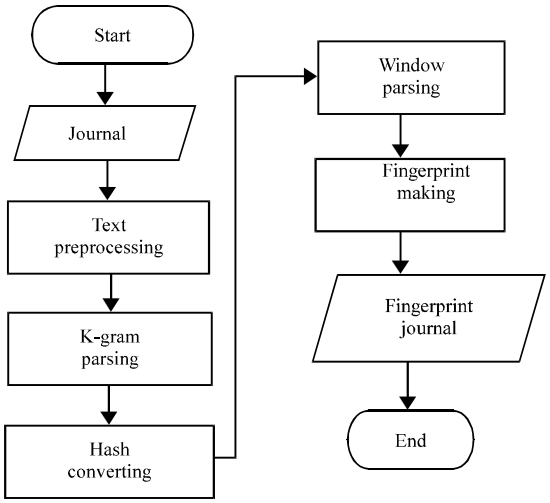

Fig. 1: Stage of the fingerprint making using algorithm

Table of field study: It is a table that saves the data of field study category which functions to divide the journal based on each category of its field study.

Table of journal: Table of journal is a table that saves journal data both specifically about the journal or the fingerprints as the result of processing using algorithm.

System design: Stage of the system designing which is going to build uses Unified Modeling Language (UML) that supports the model concept of programming based on Object Oriented Programming (OOP) as will be applied to program code writing-stage. This stage will result modeling documentation, those are business process, use case diagram, use case scenario, sequence diagram, activity diagram, class diagram and Entity Relationship Diagram (ERD).

Implementation: Implementation stage is the conversion process of the system design into the program code. A system that is going to be built is written by language program of Page Hypertext Pre-Processor (PHP) using codeIgniter framework that has applied Object Oriented Programming concept (OOP). This system also, uses local server and database to save the data that is needed anytime and can be re-accessed. The local server uses XAMPP application which supports apache to build an application based on web and database that was used MySql (PhpMyadmin).

Testing: After the implementation process, the next stage is system testing. This research conducted two methods of system testing, those are white box testing and black box testing. White box testing is testing on the module of

program coding to guarantee that the program code is clear from syntax or logical error. Black box testing is a testing that emphasizes on system functionality testing in order to get the expectation result.

The design and the planning: The design and the planning of the plagiarism detection systems using algorithm as the algorithm on similarity document search explained the functional need and also, non-functional need, use case diagram and Entity Relationship Diagram (ERD).

The functional need-system is a need that has to be owned by the system. The functional need-system of the plagiarism detection used algorithm as:

- The system which used login feature to authenticate the access right of its user
- The system can save the user's data on user register feature
- The system can update the user's data
- The system can show the user's data in the database
- The system can manage the journal data (insert, update, delete)
- The system can show the journal data of the database
- The system can show the log activity of the user
- The system is available in calculating the similarity of the journal document that is uploaded to the system
- The system is available to show the calculation result of the journal document similarity which is uploaded

The non-functional need is the addition need to complete the system. The non-functional needs are:

- System based on web
- System that uses Codelgniter framework

Usecase diagram is a diagram that is used to explain the feature of the system, it is described in the form of ellipse and user of the system is also, attached on the diagram. Use case diagram of plagiarism detector system can be seen on Fig. 2.

Entity relationship diagram is a data structure data and the relation of the data. The object in ERD is usually described in an entity which has attributes that relate to other entities. The process of system making in this research used nine entities. Those entities have been normalized to collect the data of the database that is related using one to many relations. Entity relationship diagram of plagiarism detector can be seen in Fig. 3.

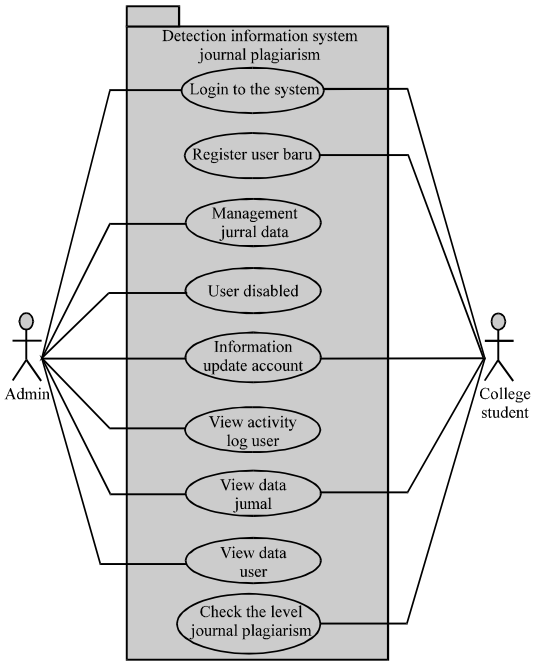

Fig. 2: Usecase diagram

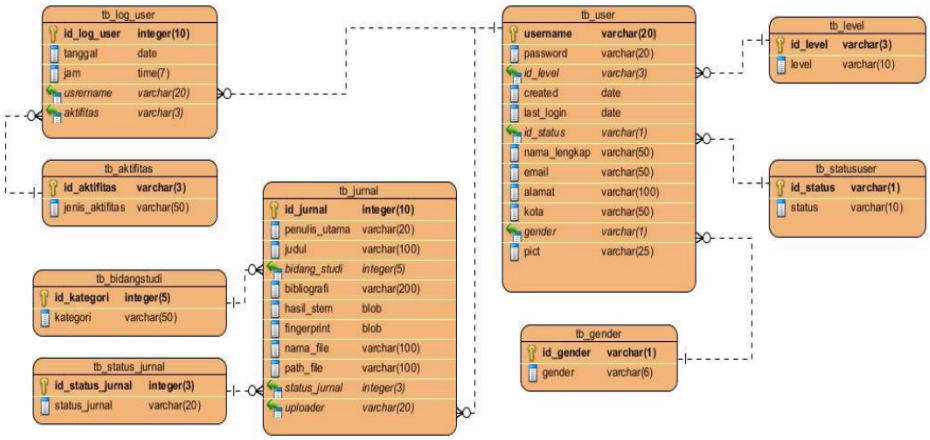

Fig. 3: Entity relationship diagram

RESULTS AND DISCUSSION

algorithm implementation in this system is in addition feature of new journal data that can be accessed by the user administration and on the feature that checks the journal plagued in which can be accessed by user students. Those two features have the fingerprint making process that uses algorithm. Every word of the journal file is uploaded first to through the filtering process that is the cleaning process of punctuation, space and etc., by referring to the basic need of algorithm detector. In this research, the writer added the stemming process that functions to change every word on the journal to be a basic word.

Stemming and filtering: Stemming is the part of retrieval (IR) information process, it changes the word into its basic word before doing the indexing process (Arifin *et al*., 2009). The stemming of Indonesian is different with the stemming of English. The stemming process of indonesian is more complicated because Indonesian has prefixes, infixes, suffixes and combination between confix.

The stemming and filtering process of the system are included in stemmer and filtering function method. Before processing the file using those methods, the file has to be processed into document extraction process. This stage is conducted in order to get the content of the document file. The process of content extraction is included into get_content () method that can be in Fig. 4.

Get_content() method needs a parameter in the form of file extension and saved file location. Furthermore, the file extension will be checked. If the file extension accomplishes the system certainty, it will be processed into extraction of the document file content. The output of this process is string in document file content which is placed in variable '$isi_file'.

The next process is stemming process and filtering process which is implemented into stemmer and filtering function method. Those methods can be seen on Fig. 5 and 6.

The stemming process on the system needs parameter input as string that is collected from the extraction process of document file content. The result of stemming process is placed in variable $hasil_stem. While the filtering process cleans the sentences from punctuation, space, the conversion of capital letters and the number elimination.

Process of k-gram forming: K-gram forming is a process that converts string into sub-string gram. k-gram is a substring series that is along with length k (Scheleimer *et al*., 2003).

The process of gram forming on the system is included in the function method of parsing_gram. Gram forming process needs two parameters. Those parameters are string of the stemming process and the filtering process and also, value k. Value k functions to check whether the result of every gram is suitable with value k that has been determined. The parsing_gram method can be on Fig. 7.

```
function get_content($ext, $file_path)
{
    $isi_file = '';
    if($ext == ".txt"){
        $isi_file = file_get_contents($file_path);
    }
    else if($ext == ".docx"){
        $striped_content = '';
        $content = '';

        if(!$file_path || !file_exists($file_path)) return false;

        $zip = zip_open($file_path);

        if (!$zip || is_numeric($zip)) return false;

        while ($zip_entry = zip_read($zip)) {

            if (zip_entry_open($zip, $zip_entry) == FALSE) continue;

            if (zip_entry_name($zip_entry) != "word/document.xml") continue;

            $content .= zip_entry_read($zip_entry, zip_entry_filesize($zip_entry));

            zip_entry_close($zip_entry);
        }// end while

        zip_close($zip);

        $content = str_replace('</w:r></w:p></w:tc><w:tc>', " ", $content);
        $content = str_replace('</w:r></w:p>', "\r\n", $content);
        $striped_content = strip_tags($content);

        $isi_file = $striped_content;
    }
    else{
        if(file_exists($file_path)) {
            if(($fh = fopen($file_path, 'r')) !== false ) {
                $headers = fread($fh, 0xA00);
                $n1 = ( ord($headers[0x21C]) - 1 );// 1 = (ord(n)*1) ; Document has from 0 to 255 characters
                $n2 = ( ( ord($headers[0x21D]) - 8 ) * 256 );// 1 = ((ord(n)-8)*256) ; Document has from 256 to 63743 characters
                $n3 = ( ( ord($headers[0x21E]) * 256 ) * 256 );// 1 = ((ord(n)*256)*256) ; Document has from 63244 to 16775423 characters
                $n4 = ( ( ( ord($headers[0x21F]) * 256 ) * 256 ) * 256 );// 1 = (((ord(n)*256)*256)*256) ; Document has from 16775424 to 4
                $textLength = ($n1 + $n2 + $n3 + $n4);// Total length of text in the document
                $extracted_plaintext = fread($fh, $textLength);
                $extracted_plaintext = mb_convert_encoding($extracted_plaintext,'UTF-8');
                $isi_file = $extracted_plaintext;
            } else {
                $isi_file = 'Cannot Load File Content';
            }
        }
        else {
            $isi_file = 'File Not Exist';
        }
    }

    return $isi_file;
}
```

Fig. 4: Get_content() -method

```php
function stemmer($isi_jurnal)
{
    // create stemmer
    // cukup dijalankan sekali saja, biasanya didaftarkan di service container
    $stemmerFactory = new \Sastrawi\Stemmer\StemmerFactory();
    $stemmer  = $stemmerFactory->createStemmer();

    // stem
    $hasil_stem    = $stemmer->stem($isi_jurnal);

    return $hasil_stem;
}
```

Fig. 5. Stemmer method

```php
function filtering($string)
{
    $filter        = preg_replace("/[^a-zA-Z]/", "", $string);
    $filter_final  = strtolower($filter);

    return $filter_final;
}
```

Fig. 6. Filtering method

```php
function parsing_gram($filter_final, $gram)
{
    $arr1 = str_split($filter_final);

    for($i=0;$i<count($arr1); $i++){
        array_slice(array_slice($arr1, $i), 0, $gram);

        if(count(array_slice(array_slice($arr1, $i), 0, $gram)) == $gram){
            $arr2[$i] = array_slice(array_slice($arr1, $i), 0, $gram);
        }
    }

    return $arr2;
}
```

Fig. 7: Parsing_gram method

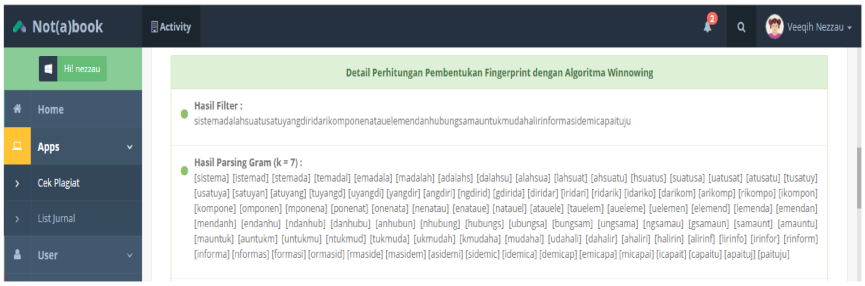

Fig. 8: The result appearance of gram forming on the system

Those are the result gram forming process showed by the system to ensure that the gram forming is suitable with the concept of algorithm. The appearance of gram forming on the comparison sentence file that has been processed on stemming can be seen on Fig. 8.

The conversion process of hash value: Firstly, every character in gram is converted into value ASCII. It is done before doing the calculation of rolling hash value. Hash formulation is defined as:

$$Hc1,..., ck = c1*b^{(k-1)}+c2*b^{(k-2)}+,...,+c(k-1)*b+c^{k} \quad (1)$$

Where:

c = Value ASCII character
b = Basis of primes
k = Value k-gram/many characters on

Rolling hash gram has an advantage to search the hash $H_{(c2...ck+1)}$ by using the earned value on hash $H_{c1, ..., ck}$ function hash $H_{(c2...ck+1}$ can be done by Eq. 2:

$$H(c2, ..., ck+1) = \left(H(cl, ..., ck) - c1*b^{(k-1)}\right)*b+c(k+1) \quad (2)$$

The forming process of the window on the system is included into the parsing_window function. Window process needs two parameters. Those are each conversion of hash value and value w. The value w functions to check whether the window forming result is suitable with value w that has been determined.

The conversion process uses the looping process for each conversion in each gram into hash value. Every character in each gran is converted into ASCII value with the help of function that has been given by language program php 'ord function'. Next, hash value is put into the array and next it is showed by the system.

The forming process of Window w: Forming window is a conversion process of hash value into windows. The window forming process of the system is included into parsing_window function method. Window forming method needs two parameters. First, it is the result of hash value and value w conversion. Value w functions to check whether the result of window forming is already suitable with value w that has been determined.

The result of window forming process is showed by the system to ensure the window forming is suitable with the concept.

The process of fingerprint selection: Fingerprint selection is the forming fingerprint process. It is conducted by selecting the minimum hash value of each different window that has been made before. If there are similar values in a different window, just choose one of those value to be the fingerprint.

The fingerprint selection process of the system is included into function fingerprint method. The fingerprint selection process needs the parameter of window w forming result. This process uses the looping function to select the minimum value on each window and array_unique function to ensure that there is no similar value on the array as the result of fingerprint selection. The fingerprint method can be seen on Fig. 9.

The result of fingerprint selection process is showed by the system to ensure that the fingerprint selection is suitable with algorithm system.

The calculation process of document similarity: The similarity calculation using Jaccard coefficient similarity. Jaccard coefficient similarity is similarity measurement standard that compares two sets, P and Q (Naumann and Melanie, 2010). The general pattern of Jaccard's coefficient similarity formulation can be seen in Eq. 3:

$$Jacc(P, Q) = \frac{P \cap Q}{P \cap Q} \quad (3)$$

Where:

Jacc (P, Q) = Similarity value
$|P \cap Q|$ = The similar fingerprints number of P and Q
$|P \cup Q|$ = Fingerprint p that does n ot have Q is added by number of fingerprint Q that does not belong to P is added by the number of similar fingerprint between P and Q

The similarity calculation process on the system is included into proses_hitung method (process_calculation). System searches the difference of the value in both array using function array_diff and then count the number of array using count function. The result of similarity calculation process can be seen in Fig. 10.

```
391     function fingerprint($parsing_window)
392     {
393         $t=0;
394         while($t<count($parsing_window)){
395             $arr5[] = min($parsing_window[$t]);
396             $t++;
397
398         }
399
400         $fingerprint = array_unique($arr5);
401
402         return $fingerprint;
403     }
404 }
```

Fig. 9: Fingerprint method

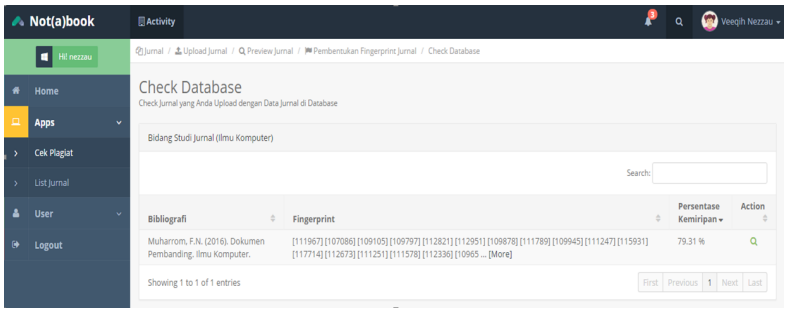

Fig. 10: The program code of similarity calculation

CONCLUSION

This application is free and can be used for every journal, daily tasks and final task that is written in Bahasa Indonesia as the priority. Users can make their own database that can check plagiarism from every journals, daily tasks and final task that is written before.

algorithm as the algorithm fingerprint document and Jaccard coefficient similarity as the similarity measurement standard can be applied in finding the similarity percentage of text document. The journal document as the input process using algorithm for the document's fingerprint-forming. The fingerprint of each document is calculated using Jaccard coefficient similarity formulation in order to get the similarity percentage. The researcher also, added the stemming process to algorithm in order to get higher accuracy.

algorithm error value with stemming process has less error value 1.63% whereas algorithm without stemming process has error value 0.69%. Hopefully, the advice below can be used for further research to reach a better result.

The addition process of undescriptive word elimination such as "yang", "dan", "di", "dari", etc., may give a higher accuracy.

The journal document can be processed by a more diverse system by adding the document in. pdf extension which has been suitable for the journal writing format (two column).

Determination of gram value an window value uses a more relevant method in order to get higher accuracy.

REFERENCES

Anonymous, 2010. [Regulation of the minister of national education]. Ministry of Education and Culture Indonesia Ministry, Jakarta, Indonesia. (In Malay)

Arifin, A.Z., I.A. Mahendra and H.T. Ciptaningtyas, 2009. Enhanced confix stripping stemmer and ants algorithm for classifying news document in Indonesian language. Proceedings of the 5th International Conference on Information and Communication Technology and System, August 4-4, 2009, Elsevier, Amsterdam, Netherlands, pp: 149-158.

Cornic, P., 2008. Software plagiarism detection using model-driven software development in eclipse platform. Master Thesis, University of Manchester, Manchester, England.

IUE., 2012. [Technical instructions of plagiat prevention]. Indonesia University of Education, Bandung, Indonesia. (In Indonesian)

Naumann, F. and H. Melanie, 2010. An Introduction to Duplicate Detection. Morgan and Claypool Publisher, Canada, ISBN:9781608452200, Pages: 51.

Oetsch, J., J. Puhrer, M. Schwengerer and H. Tompits, 2010. The system kato: Detecting cases of plagiarism for answer-set programs. Theor. Pract. Logic Programm., 10: 759-775.

Scheleimer, S., D.S. Wilkerson and A. Aiken, 2003. : Local algorithm for document fingerprinting. Proceedings of the ACM International Conference on Management of Data (SIGMOD), June 09-12, 2003, ACM Press, New York, USA., ISBN:1-58113-634-X, pp: 76-85.