A Narrative Method for Evaluating Documents Similarity based on Unique Strings

Phan Hieu Ho, Trung Hung Vo,
Ngoc Anh Thi Nguyen, Ha Huy Cuong Nguyen

Abstract--- A precision and efficiency model of the similarity computing of texts plays an important key of duplicate documents detection. In this paper, we focus on presenting and evaluating documents similarity based on a new method viaen coding text into unique strings, called Deoxyribo Nucleic Acid (DNA) sequences. Additionally, the proposed method including an algorithm for marking as well as coloring similar paragraphs in the test document compared to other documents available in the data warehouse and developing a system for copy detection are investigated. Experimental results show that this novel approach is highly accurate for areal dataset taken from PAN. The results corroborate the advantages of the novel approach with average of 99%accuracyfor the text similarity detection with a selection threshold of ε=10-12.The results of this study are applied to implement a practical system for evaluating documents similarity at the University of Danang, Vietnam.

Keywords--- Text Similarity, Text Encoding, DNA Sequencing, Text Coloring, Copy Detection.

I. INTRODUCTION

Along with the development of the Internet, the document exchanging and sharing process has become more convenient than ever. Documents such as articles, research papers, graduation theses, technical reports which used to be scarce ages ago now has become so abundant and approachable thanks to the Internet. Users can find the necessary information quickly and easily with just one click of a mouse. However, besides the advantage of providing a rich source of reference, the plagiarism is becoming a problem. The issue is how to assess the level of similarity of text to aid in the detection of text copying.

In the world, the research results on similarity assessment (in the same level) in the English text have many researches and many useful applications, including the problem copy or plagiarism detection [Bin-Habtoor & Zaher, 2012; Meuschke&Gipp, 2013; Rubini & Leela, 2013; Gomaa & Fahmy, 2013]. However, there are still many challenges that need to be solved. Although there have been many studies and systems for copy detection, there has not been a common ground in assessing their effectiveness. At the same time, research groups in this field have been started in Vietnam [Toi, Hung,& Son, 2011; Nguyen, Toan,&Dien, 2016; De, et al., 2014], and very few systems and practices have been put into practice or commercialized.

In addition, problems in natural language processing, text search and matching are documented by the local and international scientific community. For document processing tasks, text representation is an important preprocessing step. Traditional text expression models such as the bag-of-words model and vector space are the most commonly used models [Hourrane&Benlahmar, 2017; Raghavan & Wong, 1986]. In our research, we have implemented a method based on the application weighted vector model in the copy detection problem [Hung, Ngoc-Anh, Hieu, & Dung, 2017]. Although, vector-based methods have been applied to detect text copy. However, the vector representation method is still limited in the number of dimensions. Therefore, the text file will be large. As a result, the storage space increases the complexity of the algorithm when comparing and reducing the computational speed.

Thus, in this study, we have researched and proposed a new solution to solve this problem better, namely the transition from text to numbers to take the numerical advantages of large data processing, search fast and accurate. The proposed solution is based on the discrete wavelet transform (DWT) and the Haar filter [Hieu, Hung,& Ngoc Anh, 2018]. In this paper, we focus on encoding text into unique strings, called DNA sequences, empirically on the actual dataset of the PAN with the selection threshold of $\varepsilon=10^{-12}$, and the correct result is about 99%. In addition, the article also proposes an algorithm for marking, coloring the same paragraphs and developing a practical system at the University of Danang, VietNam.

II. METHODS AND EXPERIMENTAL PROCEDURE

2.1. Copy detection problem

According to Meuschke&Gipp, et al., [Meuschke&Gipp, 2013;Potthast, Eiselt, Barrón Cedeño, Stein,& Rosso, 2011], the copy detection problem can be formulated: a copy case $s = \{s_{plg}, d_{plg}, s_{src}, d_{src}\}$, where a text s_{plg} in the document d_{plg} is copied from the text s_{src} in the document d_{src} (source document). With a given document d_{plg} (the document to be examined), the task of the copy detection system is to detect s by indicating a copy case $r = \{r_{plg}, d_{plg}, r_{src}, d'_{src}\}$, it includes a paragraph that is supposed to copy r_{plg} in the document d_{plg} and its source text is r_{src} in d'_{src}, the closest approximation possible with s. They concluded that r is detectable s if and only if $s_{plg} \cap r_{plg} \neq \emptyset$, $s_{src} \cap r_{src} \neq \emptyset$ and $d_{src} = d'_{src}$. Normally, the text r_{plg} must be long enough to avoid accidental duplication, which can be set by a certain threshold t.

Manuscript received September 16, 2019.

Phan Hieu Ho, The University of Danang, Danang City, Vietnam.
Trung Hung Vo, The University of Danang, Danang City, Vietnam.
Ngoc Anh Thi Nguyen, The University of Danang- University of Education and Science, Danang City, Vietnam.
Ha Huy Cuong Nguyen, Quang Nam University, Tam Ky, Vietnam.

In the copy detection system, in order to compare the similarity between the chunks of text, s was found by searching for the document d'_{src} from a set of very large documents D and extract r_{src} and r_{plg} from the two documents d'_{src} and d_{plg} based on a detailed comparison between the two documents.

The commonality of copy detection methods is that they measure the similarity of the text T with the texts in D. The similarity measurement of two texts is usually based on measuring the similarity between paragraphs or unit elements (chunk) in T with the chunk of text in D. There are many different copy detection solutions that define different unit components. These unit elements can be words, n-grams, sentences, or paragraphs. Figure 1 shows how to handle the text copy detection.

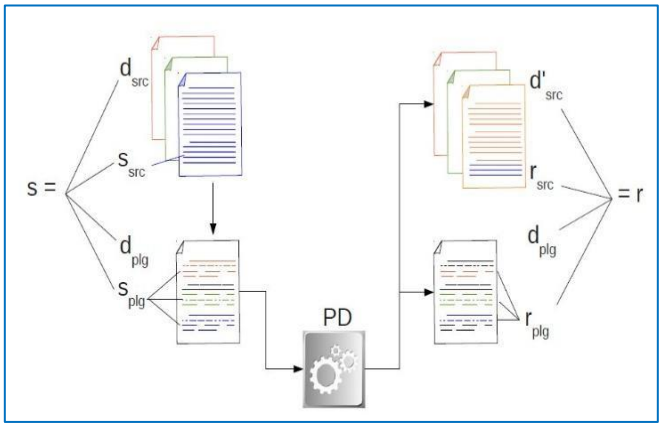

Figure 1: Describe how to handle text copy detection

2.2. DWT and Haar filters

As known, Discrete Wavelet Transformations (DWT) are simple and fast and are used very effectively in digital signal processing.

DWT signals for one-dimensional signals are described as follows: The signal is divided into two parts, the high frequency part and the low frequency part. Low frequency components are further divided into two parts of high and low frequencies; These steps are called sampling steps in down. Besides, the complexity of coding is linear and supports multiple resolutions. Multivariate analysis utilizes digital filtering techniques during analysis to filter interference and identify abnormal signals [Jeeralbhavi, Pujari, &Seeri, 2016; Mallat, 1989]. Multivariate analysis is likely as two signal filters, each of which is analyzed into two components, namely: the approximation component, A, and the detailed component, D, corresponding to the low and high frequency components, respectively. Figure 2 shows multi-resolution analysis using DWT. Two low pass filters use the proportional $\Phi(x)$ and the high pass filters use the Wavelet ψ (x) function. The relationship between the scaling function and the function Wavelet is given by:

$$\phi(x) = \sum_{k=0}^{N-1} C_k . \phi(2x - k) \quad (1)$$

$$\psi(x) = \sum_{k=0}^{N-1} (-1)^k C_k . \phi(2x + k - N - 1) \quad (2)$$

where: C_k is a scalar value for determining the scaling factor.

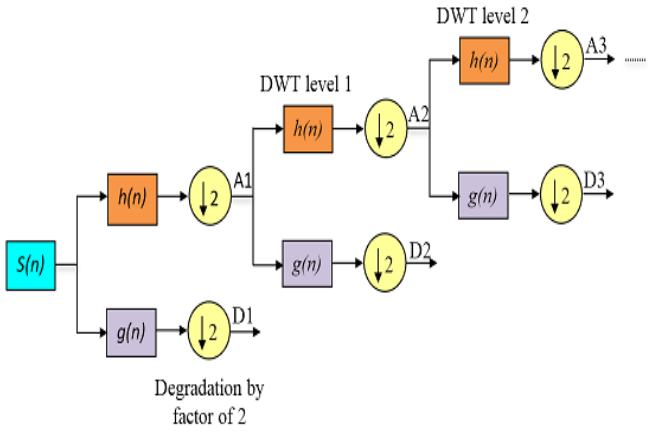

Figure 2: Multi-resolution Analysis using DWT

Filter calculations are carried out with different layers to reduce the computed mass, when passing each filter, the sampled signal is reduced twice. For each layer, the signal has a different resolution, so DWT is called a multi-resolution analysis. At each layer, the expression for the filter is given by the following formula:

$$y_{high}(n) = \sum_n S(n).g(2k - n) \quad (3)$$

$$y_{low}(n) = \sum_n S(n).h(2k - n) \quad (4)$$

where *S(n)* is the signal; *h(n)* is the impulse response of the low pass filters corresponding to the ratio function *Φ(n)*; *g(n)* is the impulse response of the high pass filter corresponding to the Wavelet function *ψ(n)*. These two filters are related in the following way:

$$h(N - 1 - n) = (-1)^n g(n) \quad (5)$$

where: *N* is the number of samples in the signal.

As mentioned above, we see DWT separating the signal into approximate and detailed components (or approximation coefficients and detailed coefficients). After decomposing the signal, reconstruct the original signal so that it does not lose information by combining the approximate and detailed components together through the inverse discrete wavelet transform. Transform - IDWT). The signal *S(n)* can be reconstructed following the inverse steps given by the formula:

$$S(n) = \sum_k y_{high}(k).g(2k - n) + y_{low}(k).h(2k - n) \quad (6)$$

Therein, $y_{high}(k)$ and $y_{low}(k)$ respectively are the output signal after passing through the high pass filter and the low pass filter.

In the case of the Haar filter, it was introduced by Hungarian mathematician Alfred Haar and put into practice in 1910 [Stanković & Falkowski, 2003]. Haar is a discontinuous wave, like a jump function. In the discrete wavelet transform, the Haar Wavelet line, also known as the Haar filter, is commonly used in data mining and indexing. The Haar Wavelet is one of the first examples of small and orthogonal wavelet transformations [Popivanov, 2001].

Haar Wavelet has two functions as a proportional function and the Wavelet function is defined by the formulas by [Chan& Fu, 1999]:

$$\psi_i^j = \psi(2^j x - i) \qquad i = 0, ..., 2^j - 1 \tag{7}$$

$$\text{with} \quad \psi(t) = \begin{cases} 1 & \text{with } 0 < t < 0.5 \\ -1 & \text{with } 0.5 < t < 1 \\ 0 & \text{with another} \end{cases} \tag{8}$$

Haar line has the form as square wave as shown in Figure. 3.

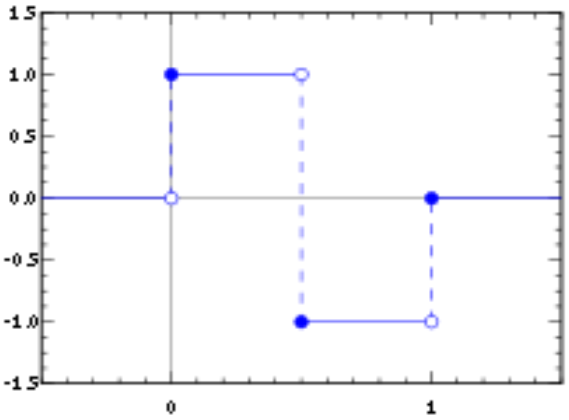

Figure 3: Haar Wavelet

Due to its low complexity, the Haar filter was primarily used for sample identification, image processing, digital signal processing,... [Jeeralbhavi, Pujari, &Seeri, 2016; Mallat, 1989].

The operating speed of the Haar filter is the fastest in all wavelet waves because the coefficient of the Haar function is 1 or -1.

Haar filters can decompose signals into different components of the frequency domain. The 1-D DWT divides the input signal into two components (that is, the average component and the component component) by calculating the low pass filter and the high pass filter [Liang & Chen, 2004].

With the input data being a sequence of numbers, Haar transforms pair two consecutive numbers in this sequence. The digits of the pairs are stored, while their sum forms a new string that is used as the input for the next conversion. This process is repeated and eventually results in a series of numbers consisting of the value of pairs of numbers over variables and a sum value. Haar transformations are a simple form of compression that includes processing such as averaging, signaling, storing details, eliminating data, and reconstructing data so that data after processing similar to the original data [Raviraj&Sanavullah, 2007].Thus, the research using the Haar filter to convert real-time serial signals into DNA for calculating, treating and filtering signals is feasible and highly feasible.

2.3. Proposed method

Through research on DWT and Haar wavelet, we propose the idea to convert the text into a sequence in real time (via digitizer) and use filters Haar in DWT to detect patterns weirdo, the text data to be converted and expressed as ranges represented by real numbers $\mathbf{x} = [x_1\ x_2 \ldots x_N]$. In the following, the proposed process of coding the text into a real DNA sequence and the algorithm for coloring identical text for use in a text-to-speech detection system will be analyzed.

2.3.1. The process of encoding text into real numbers

Text data is converted and expressed as real numbers (T = t_1, t_2, ..., t_n). To determine the degree of similarity between the two sequences X = x_1, x_2,…, $x_n$andY = y_1, y_2,…, y_n need to calculate the Sim (X, Y) similarity of these two real numbers. We recommend the process of encrypting text into digital signals, as shown in Figure. 4.

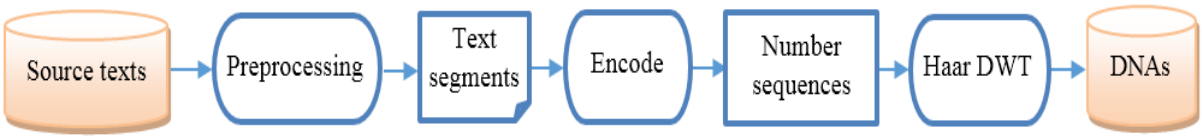

Figure 4: The process of encoding text into real numbers

In this study, we focus on designing blocks to encode textual data into DNA that serves to evaluate the similarity of text. Table 1 shows the process of encoding text into digital signals.

Indeed, the available materials are collected first, and preprocessing removes punctuation, special characters, and archives as raw data. In order to facilitate the main processing, during the pre-processing phase, the collected text will be segmented and sampled so that the samples are of equal length.

These segments are then stored as raw data for extracting the same paragraphs (if any) at the output of the evaluation results.

Finally, in the main processing stage, the text will be digitized into a sequence of numbers and passed through the Haar filter to obtain data for the DNA sequence.

Algorithm 1: The process of encoding text into digital signals

Input: Document
Output: DNA sequences
Process: Encode text into digital sequence
Step 1: Preprocessing (removing punctuation, special characters, indexing and raw data storage, etc.)
Step 2: Digitize to convert raw data into serial numbers
Step 3: Process through Haar filter to encode into DNA sequences

2.3.2. Organization of the data for the source DNA kit and similarity detection

After performing the steps in the digitization process, a set of DNA is obtained, corresponding to the set of collected documents.

The DNA at the first value of the DNA incrementally is sorted. This arrangement allows the system to perform a binary search to identify whether the sorted DNA is similar to the DNA of a particular segment of the text. Thereby, the complexity of the text evaluation algorithm can be improved. It is possible to use the first value of DNA as a sort key since it is the approximate or sum value of the component values after the iteration step. So, at this point, if the values of the two DNA samples (a sample of the source text and an evaluation text sample) are the same, two text samples corresponding to these two DNAs will be identical.

The second step is to compare each group of DNA sequences the stored DNA of the source dataset. For each sample of DNA in the DNA group to be included in the comparison, we will look for binary in the data warehouse to determine which source DNA is at the first similarity to the most likely DNA. Next, the Euclidean distance between the two DNAs is calculated very simply by the following formula:

$$d(\mathbf{x},\mathbf{y}) = \|\mathbf{x}-\mathbf{y}\|_2^2 \quad (9)$$

Therein, $\mathbf{x} \in \square^{1\times N}$ and $\mathbf{y} \in \square^{1\times N}$ respectively as the source DNA vector and the DNA vector under consideration. This Euclidean distance will be compared to a threshold level ε. If $d(\mathbf{x}, \mathbf{y}) < \varepsilon$, the two DNAs are considered to be identical and the position corresponding to the DNA under consideration is re-marked for the decision-making system after synthesizing all the DNA samples from the segment.

In this proposed approach, we have also considered large data processing with the encoding of text data into digital form, sorting the data in ascending order that allows binary search, as this is One of the fastest search methods when working with large data. In addition, DWT for computational complexity is only a linear function in each subset of sampling, so the proposed solution will be more effective in large data processing.

2.3.3. Generic text colors algorithm

Coloring the copy texts in the document file is necessary for performance evaluation, as well as showing the similarity between two texts. The text coloring, for instance, is to perform a match between the Word input file (.doc, .docx) and the documents in the system's data warehouse, after converting the text data into DNA. The same sentences and paragraphs will be implemented as the following model (see Figure 5).

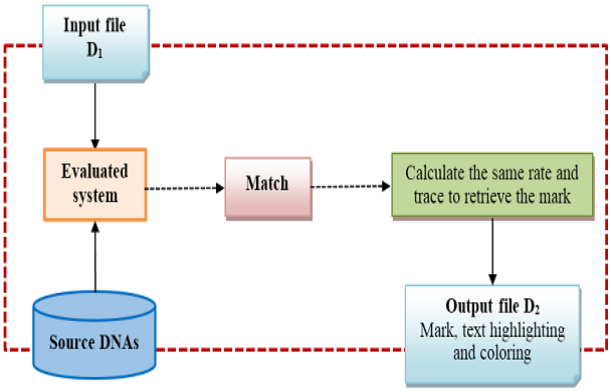

Figure 5: The model detection pattern marking the same content

The steps to detect and mark the contents of the text that similar each other are introduced, as follows:

Step 1: To conduct a match between the input file (.doc or .docx) and the data warehouse (source DNAs) of the system, depending on the field and specialty of the input text, the user can select the field corresponding to zoning, limiting the number of documents in the database when matched makes searching faster.

Step 2: Make a match of the DNA fragment of the document to be examined (input file) with the source DNA stored, indexed in the sorted database.

Step 3: Calculate the same rate and trace to retrieve the mark. After checking the DNA fragments of the text that needs to be checked against the source DNA stored in the system, the tool will keep track of the same paragraphs for access in the source documents.

Step 4: From the track of the sentence, the paragraph is the same from the source of other documents in the system, the tool will mark the comment, the paragraph is the same document, same year and same rate of the above.

Step 5: Mark and color. Depending on the level of the sentence, the system will change the format and color the text with the following levels:

Step 5.1: Similarity from 90% to 100%: red;

Step 5.2: Similarity from 70% to 89%: blue;

Step 5.3: Similarity from 50% to 69%: yellow.

Text highlighting and coloring algorithms are implemented as shown in Algorithm 2.

Algorithm 2: Mark and color the similar paragraphs

Input: Text (.doc or .docx file)
Output: Text highlighted, highlighted copying suspects, and references to copied source documents
Process:
Step 1: n = CountSent(D_1) // *Number of statements of the file to be tested D_1*
Step 2: **For***i*: = 1 → *n*
Step 2.1: m = length(W) // *Number of sentences in the data store*
Step 2.2: Extract S_i // *Split the ith sentence in D_1*
Step 2.3: Encode S_i // *Encodes the ith sentence in D_1 into DNA*
Step 2.4: **For***j*: = 1 → *m*
Step 2.4.1: $S_j = DNA_j$ // *DNA of the jth sentence in W*
Step 2.4.2: **If** Match(S*i*, S*j*) overlap (90% -100%): Insert note, fill in red
Step 2.4.3: **If**Match(S*i*, S*j*) overlap (70% -89%): Insert note, fill in blue
Step 2.4.4: **If**Match(S*i*, S*j*) overlap (50% -69%): Insert note, fill in yellow
Step 2.5: **EndFor** // *End of loop for line Step 2.4*
Step 3: **EndFor** // *End of loop for line Step 2.2*

III. EXPERIMENTAL RESULTS AND DISCUSSIONS

3.1. Experiment on the PAN dataset

In this paper, the PAN training data is used for implementing[PAN dataset, 2009]. The experimental data is conducted from the PAN training data in 2009 for each of the 100 suspected documents completely different from the data in the data warehouse via choosing the threshold value $\varepsilon = 10^{-12}$with achieved performance ofprecision (*prec*) and recall (*rec*) ratio over 99%.Experimentation on the PAN standard dataset is used by many research groups and laboratories around the world to evaluate methods of copy detection as well as the use of measurements for evaluation in field experiments. The experimental results using our proposed method for over 10 experiments are shown in Table 1.

Table 1: Experimental results with threshold $\varepsilon = 10^{-12}$

Test order	|S|	|D|	*prec*(%)	*rec* (%)
1	13635	13543	99.94	99.27
2	12968	12880	99.94	99.26
3	21040	20950	99.91	99.49
4	15067	14916	99.93	98.93
5	16771	16637	99.89	99.09
6	17983	17930	99.91	99.62
7	15684	15582	99.91	99.26
8	14513	14422	99.94	99.32
9	14245	14151	99.89	99.23
10	14860	14737	99.95	99.13
Average	***15676.6***	***15574.8***	***99.92***	***99.26***

Indeed, as shown in Table 1, it is found that the proposed algorithm provides very high accuracy, given by the values of *prec* and *rec*. Figure 6 indicates the experimental results with threshold $\varepsilon = 10^{-12}$using our proposed method. The PAN results show that the results are completely reliable to evaluate algorithms, new approaches as well as algorithm proposed by us.

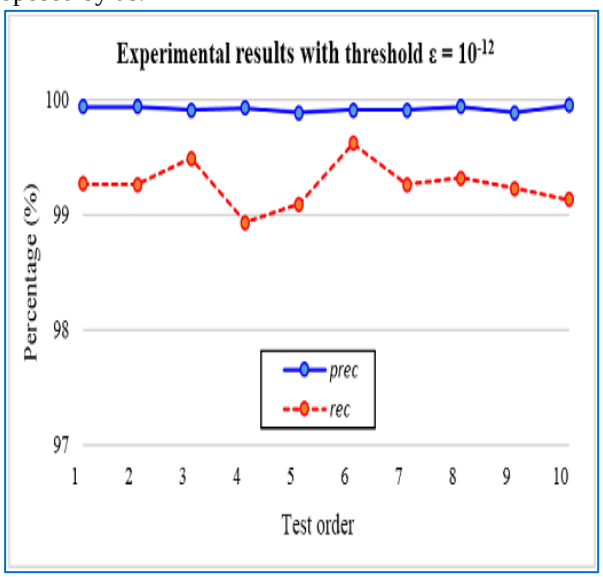

Figure 6: Experimental results with threshold $\varepsilon = 10^{-12}$

3.2. Implement the copy detection system

In order to prove that the proposed method is significant to use in practice, a system for detecting copying and experimenting at the University of Danang, VietNam is proposed. Figure. 7 shows the interface of the copy detection.

Figure 7: User Interface

An example of the same text will be highlighted, and annotated as shown in Figure. 8.

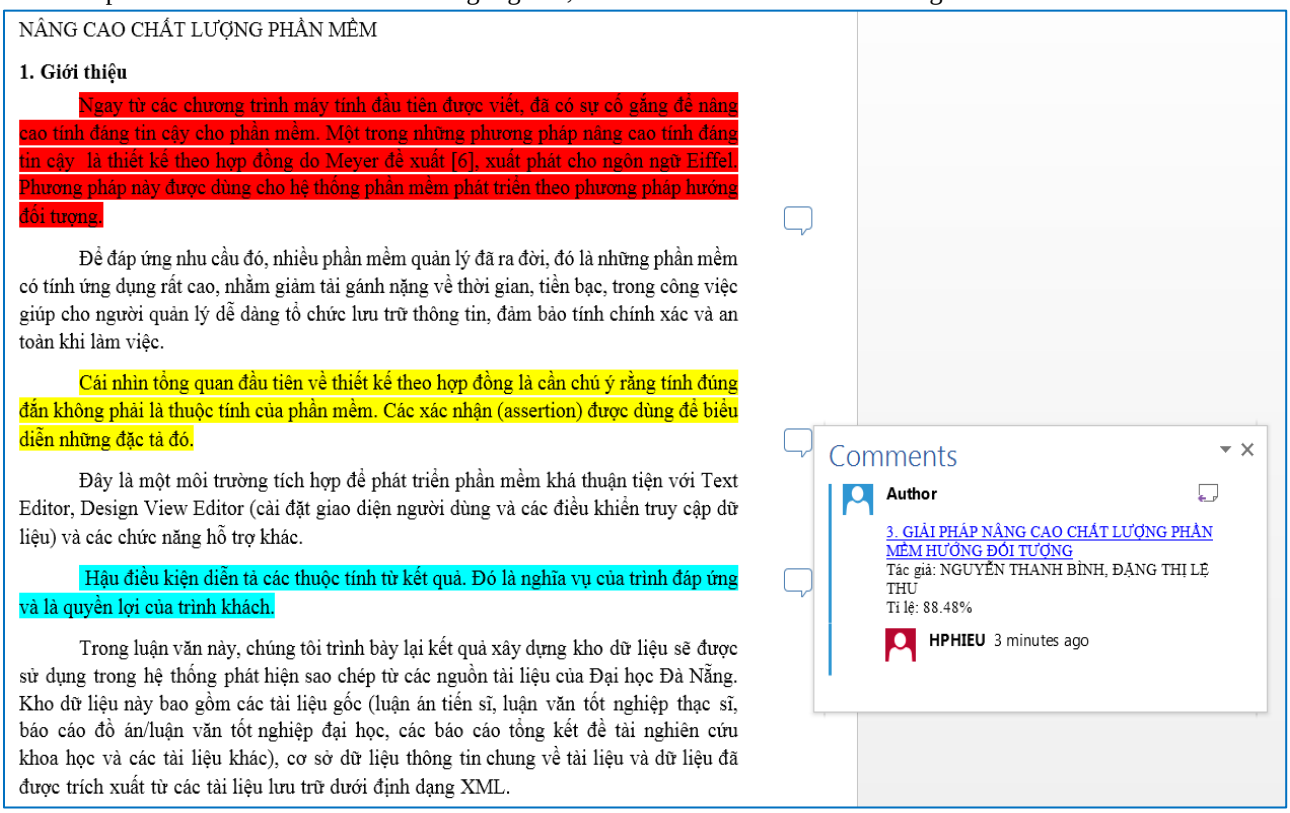

Figure 8: Example of marking and coloring the same content on the document need to be tested

The results of comparison and coloring the same paragraphs in the test document against other documents in the data warehouse. The red text is almost identical (90% - 100%), the green segment is very similar (70% - 89%). Clicking on the colored text shows the document's name, author, and the same percentage (%) of the document in the repository.

Thus, after putting a document into the copy detection system, based on the proposed algorithm, the system will return a document with their paragraphs (if any) compared, colored, and marked as similar to the paragraphs of texts in the documents stored in the data warehouse.

The system has met the requirements set and the results of the same paragraph detection with high accuracy. At present, there are nearly 2,000 documents in the assessment database, including doctoral dissertations, master theses, graduate student theses, scientific reports, scientific papers, etc. in many different fields. In the coming time, we will continue to update the new documents to the warehouse for higher coverage.

IV. CONCLUSION

In this paper, a novel method for converting text into DNA sequences based on the DWT method and the Haar filter is proposed. Then, the proposed method is used for PAN data sets for evaluating. The similarity document detection shows that the proposed method is highly accurate for the dataset. In addition, we have proposed algorithms for marking, highlighting identical text snippets in the evaluation document and pointing out the copied documents in the data warehouse. The experiments are conducted on the real dataset of PAN with a selection threshold of $\varepsilon=10^{-12}$, and getting the result is about 99% correct. Furthermore, the experimental results of this study are implemented for evaluating documents similarity at the University of Danang in practice.

REFERENCES

1. Aisyah, S. (2015). Computer networking company in business area. International Research Journal of Management, IT and Social Sciences, 2(7), 1-4. Retrieved from https://sloap.org/journals/index.php/irjmis/article/view/311
2. Bin-Habtoor, A. S., & Zaher, M. A. (2012). A Survey on Plagiarism Detection Systems. *International Journal of Computer Theory and Engineering*, *4*(2), 185-188.
3. Chan, K. P., & Fu, W. C. (1999). Efficient time series matching by wavelets. *Proceedings of the 15th IEEE International Conference on Data Engineering*, (pp.126-133). Sydney.
4. De, T. C., et al. (2014). Developing Plagiarism Detection System for Vietnamese University. *12th Vietnam - Japan International Joint Symposium*. Can Tho.
5. Gomaa, W. H., & Fahmy, A. A. (2013). A survey of text similarity approaches. *International Journal of Computer Applications*, 68(13), 13-18.
6. Hieu, H. P., Hung, V. T., & Ngoc Anh, N. T. (2018). DNA Sequences Representation Derived from Discrete Wavelet Transformation for Text Similarity Recognition. In H. Q. Thuy (Ed.), *Modern Approaches for Intelligent Information and Database Systems* (pp. 75-85). Springer SCI Book.
7. Hourrane, O., &Benlahmar, E. H. (2017). Survey of Plagiarism Detection Approaches and Big data Techniques related to Plagiarism Candidate Retrieval. *Proceedings of the 2nd international Conference on Big Data, Cloud and Applications*, (pp.

15.1-15.6). Tetouan, Morocco: BDCA.
8. Hung, V. T., Ngoc-Anh, N. T., Hieu, H. P., & Dung, D. T. (2017). Comparison of the Documents Based On Vector Model: A Case Study of Vietnamese Documents. *American Journal of Engineering Research (AJER)*, *6*(7), 251-256.
9. Jeeralbhavi, T. M., Pujari, J. D., &Seeri, S. V. (2016). Text Extraction and Localization from Captured Images. *International Journal on Recent and Innovation Trends in Computing and Communication (IJRITCC)*, *4*(6), 119-121.
10. Liang, C. W., & Chen, P. Y. (2004). DWT based text localization. *International Journal of Applied Science and Engineering*, *2*(1), 105-116.
11. Mallat, S. G. (1989). A theory for multiresolution signal decomposition: the wavelet representation. *IEEE transactions on pattern analysis and machine intelligence*, *11*(7), 674-693.
12. Meuschke, N., &Gipp, B. (2013). State-of-the-art in detecting academic plagiarism. *International Journal for Educational Integrity*, *9*(1), 50-71.
13. Nguyen, L. T., Toan, N. X., &Dien, D. (2016). Vietnamese plagiarism detection method. *Proceedings of the Seventh Symposium on Information and Communication Technology*, (pp. 44-51). Ho Chi Minh City, Vietnam: SoICT.
14. PAN dataset (2009). Retrieved fromhttp://www.uni-weimar.de/medien/webis/corpora/corpus-pan-labs-09-today/pan-09/pan09-data/pan09-external-plagiarism-detection-training-corpus-2009-03-30.zip
15. Popivanov, I. (2001). *Efficient similarity queries over time series data using wavelets*. National Library of Canada.
16. Potthast, M., Eiselt, A., Barrón Cedeño, L. A., Stein, B., & Rosso, P. (2011). Overview of the 3rd International Competition on Plagiarism Detection. In CEUR Workshop Proceedings.
17. Raghavan, V. V., & Wong, S. M. (1986). A critical analysis of vector space model for information retrieval. *Journal of the American Society for information Science*, *37*(5), 279-287.
18. Raviraj, P., &Sanavullah, M. Y. (2007). The modified 2D-Haar Wavelet Transformation in image compression. *Middle-East Journal of Scientific Research*, *2*(2), 73-78.
19. Rubini, P., & Leela, M. S. (2013). A Survey on Plagiarism Detection in Text Mining. *International Journal of Research in Computer Applications and Robotics*,*1*(9), 117-119.
20. Stanković, R. S., &Falkowski, B. J. (2003). The Haar wavelet transform: its status and achievements. *Computers & Electrical Engineering*, *29*(1), 25-44.
21. Toi, N. X., Hung, N. V., & Son, P. B. (2011). A unified plagiarism detection framework. *VNU Journal of Science: Mathematics-Physics*, *27*(1), 55-62.
22. Xavier, I. M. D. D. G. (2015). Email issue for working at information technology field. International Research Journal of Management, IT and Social Sciences, 2(5), 1-5.

Phan HieuHo received the B.S. degree in Information Technology from University of Science and Technology, the University of Danang, VietNam, in 2003, the M.Sc. degree in Computer Science from the University of Danang, in 2009. He is currently a Ph.D. student in Computer Science at the University of Danang. His research interests include Data mining, Machine learning, Text matching. Email: hophanhieu@ac.udn.vn

Trung Hung Vo is a Professor of Computer Science at the University of Danang, Vietnam. He is the Vice-Rector (Research and International Collaboration) University of Technology and Education, the University of Danang. He received his Ph.D. in Computer Science from Institut National Polytechnique de Grenoble (INPG), France, in 2004. His current research interests include Natural Language Processing, Automatic translation, Multilingual softare. Email: vthung@ute.udn.vn

Ngoc Anh Thi Nguyen received the B.S. degree in Information Technology from University of Education, the University of Danang, VietNam, in 2006, the M.Sc. degree in Computer Science in 2011, and the Ph.D. degree in Computer Science in 2016 from the Chonnam National University, Korea. She now is lecturer at University of Education, the University of Danang, Vietnam. Her current research interests include Data mining, Machine learning, Pattern recognition, Mathematical modeling. Email: ngocanhnt@ued.udn.vn