The Judgment of Document Similarities Orthogonal Transformations and Improvement of the property

Atraru Matsuzawa, Masakazu Higuchi, Gamba Jonah, Shuji Kawasaki, and Hitomi Murakami

Abstract— The objective of this work is to propose a new method to evaluate similarity of documents using Orthogonal transformations and examine its performance experimentally expecting its application to plagiarism detection. The numbers are transformed by the Fourier, Cosine, Haar Wavelet and Hadamard system. As a measure of the similarity of two documents, the correlation of them is used. As a result of the experiment, it turns out that Fourier transformation is effective for certain similar documents presumed to be produced by plagiarism. Therefore we focus on Fourier system. In case single-byte characters and double-byte characters is intermixed there is a problem. That is result of document similarities is low thought documents are similar. Against this problem we propose character-code-conversion. To standardize double-byte characters by character-code-conversion, the precision of judge rise. Furthermore, we decided threshold of each length of sentences.

Keywords—Document similarity, Orthogonal transformation, Character code, Plagiarism, Correlation coefficient, Character-code-conversion

I. Introduction

AS indicated by the word Information Big Bang, the information carried on internet has become enormous. Sometimes the information may be redundant or duplicate so that it reduces a quality of information and causes an inconvenience for users. Thus it is basically important to identify and remove documents of such undesired information. More recently, plagiarism through internet, by using just "copy and paste", has increased. Those who are to make reports and submit them in certain official way, despite that they are supposed to deliberate about the theme of the reports by themselves, can copy others' ideas collectable on the internet easily. In this sense as well, the importance of finding similar documents has further increased.

This paper proposes a new method to measure the similarities of documents. For an experimental study, we examine characteristics and effectiveness of the proposed method. In such fields as speech recognition, text search, etc., various methods based on morphological analysis are known. In the morphological analysis, a sentence is decomposed into parts of speech and then its composition is compared with database that is in a dictionary.

In our method, we take the orthogonal transformations of a sequence of character codes corresponding to a document in question, as in [1]. Then characteristics of the document are extracted by a similarity analysis on the transformed domains. Our method is simple but independent of the dictionaries and languages, which may be considered to be useful in coping with the Information Big Bang.

Manuscript submitted October 27 11, 2011. Ataru Matsuzawa is with the Graduate School of Science and Tecnology, Seikei University, Japan (e-mail: dm106225@cc.seikei.ac.jp).

Masakazu Higuchi and Hitomi Murakami are with Depertment of Computer and Information Science, the Faculty of Science and Tecnology, Seikei University, Japan (hi-murakami@st.seikei.ac.jp).

II. About character codes

There are several types of character codes. Those used for Japanese are the Shift-JIS, EUC-JP, JIS and UTF-8. In addition, the ASCII code is used for the English alphabet, while the Unicode for multiple languages. A character code is assigned to an integer on the computer. Single-byte characters are represented by 1 byte characters while double-byte characters are represented by 2 byte characters.

Depending on the character code, the alphabet, numbers, and Japanese Katakana characters are handled as single- or double-byte characters. Japanese Hiragana and Kanji are handled as double-byte characters. Taking the integer data stream as an input signal, orthogonal transformations are performed and the characteristics in the transformed domains are investigated in order to determine similarity.

The character codes support both single-byte and double-byte characters. Thus, to enable experiments for documents containing both Japanese and English as is the case of today's Japanese documents, the Shift-JIS code is considered to be suitable for this study, and we do so.

III. Orthogonal transform

In this paper, for the transformation of documents, the four representative transforms widely used in digital processing, i.e., the Discrete Fourier Transform (DFT), the Discrete Cosine Transform, the Ordered Fast Haar Wavelet Transform (HWT) and the Hadamard Transform [2][3][4], can be represented by the following formulas:

$$F_j = \sum_{k=0}^{n-1} f_k e^{-\frac{2\pi i}{n} jk} \quad (1)$$

$$F_{cj} = \sum_{k=0}^{n-1} f_k \cos\{ \frac{\pi}{n}(k + \frac{1}{2}) j\} \quad (2)$$

Equations (1) and (2) are the DFT and the DCT formulas respectively. In the above equations, f_k is the input signal data sequence, F_j, F_{cj} are the DFT and DCT coefficients, respectively. n is the number of input data points and i is the imaginary unit. The result of transforming the input signal data by Equation (1) is complex number, so for similarity comparison we use the amplitude spectrum.

$$f'_j = \begin{cases} \frac{f_{2j} + f_{2j+1}}{2} \ (0 \le j \le \frac{N}{2} - 1) \\ \frac{f_{2j} - f_{2j+1}}{2} \ (\frac{N}{2} \le j \le N - 1) \end{cases} \quad (3)$$

Equation (3) shows Haar Wavelet Transform. N is the number of data points and is a power of two.

$$F_h = \frac{1}{\sqrt{N}} Hf \quad (4)$$

Equation (4) shows the Hadamard Transform. F_h denotes the Hadamard coefficients, while f is the input signal. F_h and f are N-dimensional vectors. H is an $N \times N$ Hadamard matrix. For N=4, the Hadamard matrix is shown in Equation (5).

$$H = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \end{bmatrix} \quad (5)$$

IV. Experimental Model for Plagiarized Document

A. How to Produce the Plagiarized Document

What is examined in the experiment is that for a document in question if the proposed method of searching a document of high similarity works well. We evaluate the result as well. To this end, we will prepare an "artificially" produced document of high similarity which is supposed to be plagiarized.

As the plagiarized documents, we produce them according to the following rules:

- Replace a part of the original document by a word of 0-5 Character (hereinafter referred to as "Replacement")
- Shuffle some characters in the document ("Shuffle")
- Insert a word in the original document ("Insertion")

Although other rules, such as interchanging Kanji and Hiragana or substituting by a similar phrase that is different just in expression, may be adopted as well, we focus our investigation of similarity analysis here on those documents produced by above rules ①-③.

B. Experimental Method for Plagiarized Document Model

In this study, we conducted experiments using five different documents with the four transforms. Since for the Hadamard Transform and Haar Wavelet Transform we impose the condition that the input data points must be a power of two and that the data points must be close to the number of characters in the plagiarized document, we set the number input characters to 128. Below is a sample 128-character document that is for the "Replacement" experiment.

Example：「文章間の類似性を判定するにあたって、辞書に依存せず判断することを試みる。対象とする文字はコンピュータ上の文字で、各文字には文字コードが割り当てられている。文字コードは整数で割り当てられていて、半角は１つ全角は２つの数字が割り当てられている。例文１作成終了、」

In Experiment 1, in order to evaluate the degree of relative similarity of documents, two unplagiarized natural documents (hereinafter referred to as non-similar documents) were used. Fifteen kinds of phrases and sentences endings with different genres (music, television, sports, etc.) were prepared, from which 2 character coded ones were selected, and a total of 105 sets of cross-corrections were computed. In this experiment, cross-correlations for each of the three cases, i.e., double-byte characters only, single-byte characters only and mixed-byte characters, were calculated. In Experiments 2 to 4, two similar documents [5] were evaluated. The similar document examples are shown below.

Example of similar documents：

A: (Original) 山川惣治はCBAの説に沿って、空飛ぶ円盤を以下のように考えていました。
(Plagiarized) 山川自身はＣＢＡの説に沿って、空飛ぶ円盤を以下のように考えていたらしい。

B: (Original) 彼が初めて円盤を目撃したのはその２日後のことです。
(Plagiarized) しかし、その2日後、彼もやっと、円盤の目撃を実現させる。

The details of the experiments are as follows. Experiment 2 is on “Replacement”, Experiment 3 on “Shuffle” and Experiment 4 is on “Insertion”. The input data is the original data and the modified one and on this data the four different orthogonal transforms outlined above are applied. Based on analysis of Japanese plagiarism, the majority of the cases are in double-bye characters, and for that reason we only performed experiments on double-byte characters in Experiments 2 to 4.

In Experiment 2, we replace 1 to 5 characters in the original document and then calculate the cross-correlation between the original data and the modified data. In this case, the 1 to 5 characters are inserted to the left the document and then shifted to the right. Fig. 1 shows an example in which two characters are shifted. The right part of the figure shows the modified document.

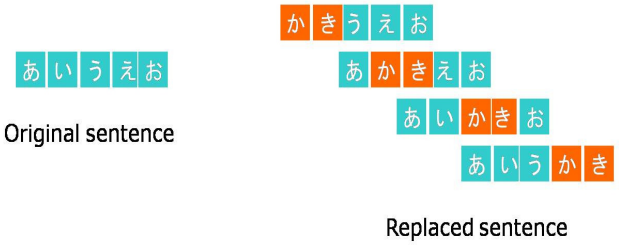

Fig. 1 an example of "Replacement"

In Experiment 3, we perform the "Shuffle" experiment. 1 to 5 characters of the original data are shifted and then the correlation coefficients between the original and the modified data are calculated. Fig. 2 shows an example in which two characters are shifted. The right side of the figure shows the modified document.

Fig. 2: an example of "Shuffle"

In Experiment 4, we perform the "Insertion" experiment. On inserting the different words into the original document, the text is only shifted by the inserted characters. By paying attention to this shift, the autocorrelation of the non-overlapping 128 characters of the original document was taken. In this experiment, by increasing the number of shifts, the compared documents increasingly become non-similar. By shifting the text, the number of characters to be compared continually decreases. In this experiment, the shift was repeated until only one character remained. Fig. 3 shows the case of 5 characters.

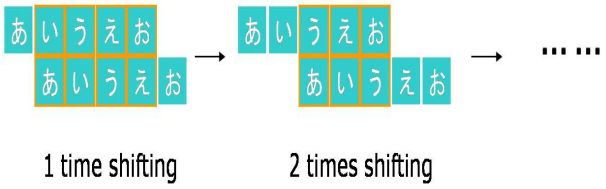

Fig. 3 an example of "Insertion"

V. Experimental Results

A. Experiment 1: (Non-similar documents) Result

We prepared the following three samples as the dissimilar documents for the Experiment 1. The results of taking correlation coefficients of the original and plagiarized documents themselves, before applying the orthogonal transformations, are shown in Table 1. The result here will serve as a basis of the desired Experiments 2 to 5 below, in that we know how much correlation dissimilar documents have; This will help us, when inspecting plagiarized documents, understanding if the correlations are non-trivial or not. The results for Experiment 1 are shown in Tables 1 to 5. The non-similar documents used, i.e., double-byte only (A and B) and mixed-byte (C) are shown below. As a representative result of natural documents, the results for the mixed double-byte and single-byte characters data are shown in Fig. 4 to 8. These are the plots of the correlation coefficients determined from the combination of 105 sets.

Example of non-similar documents (double-byte characters only)

A:「この後も、東北の日本海側は曇りや雨でしょう。雷が鳴ったり、雨脚の強まる恐れがあります。そのほかは晴れる所が多いですが、大気の状態が不安定となり、所々でにわか雨や雷雨がありそうです。局地的に激しく降ることもありますので、急な強い雨にご注意ください。最高気温は」

B:「同じＣ言語を使ったプログラムでも、コンソールアプリケーションとＷＩＮＤＯＷＳ（ＧＵＩ）アプリケーションでは、作り方が全く異なります。本書は、ＷＩＮＤＯＷＳ（ＧＵＩ）アプリケーションの作り方を初歩の初歩から解説したものです。ただ、前提としてＣ言語そのものは知」

Example of non-similar documents (mixed-byte characters)

C:「2010サッカーFIFAワールドカップ・テレビ放送予定。6/11-7/11、一ヶ月にわたってグループリーグ48試合、決勝トーナメント16試合、計64試合の激闘が繰り広げられる第19回ワールドカップ。地上波テレビ放送、BSデジタル放送、CS放送（スカパー！）で生中継される全64試合の日程・」

TABLE Ⅰ the average/standard deviation of the correlation coefficients of non-similar documents with no transform applied

	Double-byte	Single-byte	Mixed-byte
Average	0.471	0.030	0.030
Std. Dev.	0.086	0.092	0.186

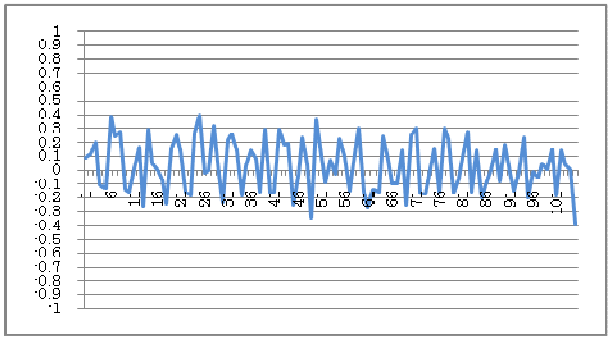

Fig. 4 correlation coefficients of mixed single-byte and double-byte documents for no transformation

When an ordinary character coded document is used as the input signal, the correlation coefficients are shown in Fig. 4. It can be concluded from Table 1 that the standard deviation of the mixed-byte documents is high.

TABLE II. average/standard deviation of the correlation coefficients for Experiment 1 input signal by the DFT

	Double-byte	Single-byte	Mixed-byte
Average	0.890	0.995	0.751
Std. Dev.	0.052	0.004	0.076

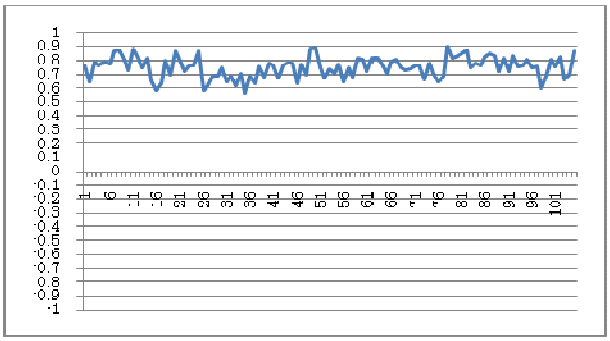

Fig. 5 correlation coefficients of mixed single-byte and double-byte documents by the DFT

On comparing Table 1 with Table 2 the frequency domain correlation coefficients by the DFT are high when compared to those taken without any transformation. The double-byte document has a high correlation value of about 0.9.

From Tables 1 and 2, it can be observed that the standard deviation of correlation coefficients after DFT is low for all document samples. From this fact, it can be said that the DFT of coded documents results in the narrowing of the range of correlation coefficients.

TABLE III. the average/standard deviation of the correlation coefficients for Experiment 1 input signal by the DCT

	Double-byte	Single-byte	Mixed-byte
Average	0.788	0.988	0.548
Std. Dev	0.055	0.009	0.118

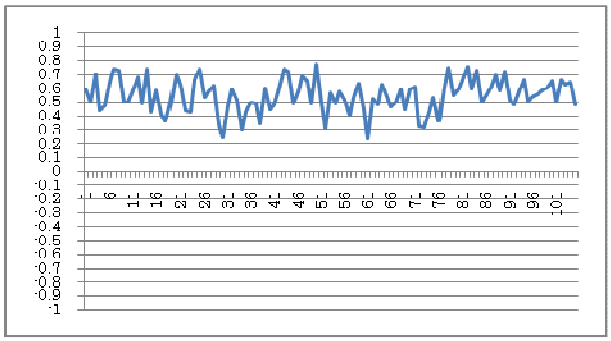

Fig. 6 correlation coefficients of mixed single-byte and double-byte documents by the DCT

In Tables 2 and 3, the average of correlation of DFT image is larger than that of DCT, while the standard deviation smaller. This implies that the magnitude and concentration of correlation of DFT images are higher than those of DCT images.

TABLE IX. the average/standard deviation of the correlation coefficients for Experiment 1 input signal by the HWT

	Double-byte	Single-byte	Mixed-byte
Average	0.409	0.585	0.038
Std. Dev.	0.088	0.145	0.210

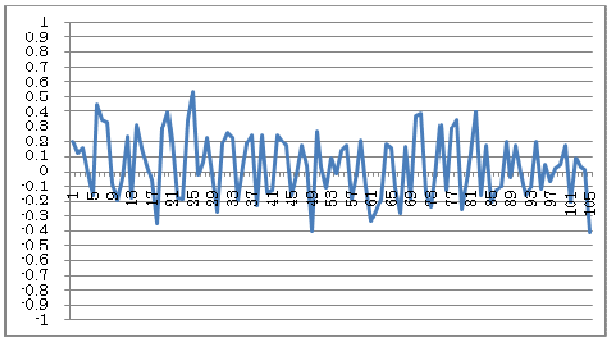

Fig. 7 correlation coefficients of mixed single-byte and double-byte documents by the HWT

From Tables 1 and 4, the HWT results, when compared with the result in which no transformation is applied, has large values of STD. When compared to the DFT and the DCT, the average values are low; the correlation of HWT is thus similar as those no transformations performed.

TABLE X. the average/standard deviation of the correlation coefficients for Experiment 1 input signal by the HT

	Double-byte	Single-byte	Mixed-byte
Average	0.701	0.977	0.391
Std. Dev.	0.067	0.017	0.139

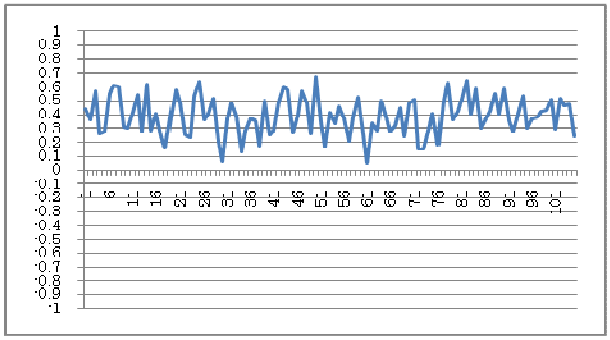

Fig. 8 correlation coefficients of mixed single-byte and double-byte documents by the HT

On comparing Tables 5 with Tables 3 and 4, it can be observed that the average and the STD values of the HT are within those of the DFT and the DCT. For single-byte characters, the values are as high as those of the DFT and the DCT.

B. Experiment 2: ("Replacement" ①) Result

In Experiment 2, a change of n characters, $1 \leqq n \leqq 5$, in the original document led to comparatively no significant difference in the result. The horizontal axis represents the "time"-lags for the modified characters.

From Fig. 9 to 13, it can be observed that for the 128 character document, a modification of characters with n around 5 results in a high correlation of about 0.9 irrespective of whether transformation is applied or not. Therefore, for this level of modification, it may be said that whether we apply the orthogonal transformations or not do not cause significant differences.

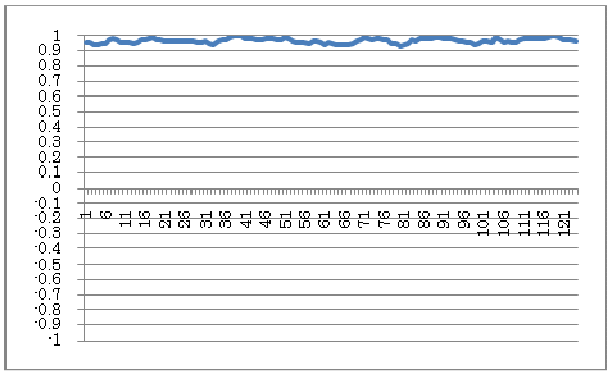

Fig. 9 experimental result for "Replacement" (no transform)

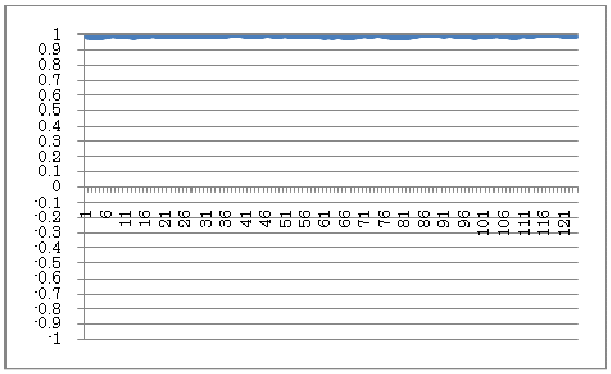

Fig. 10 experimental result for "Replacement" (DFT)

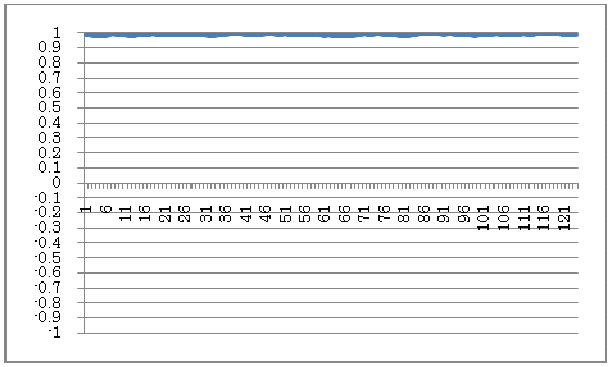

Fig. 11 experimental result for "Replacement" (DCT)

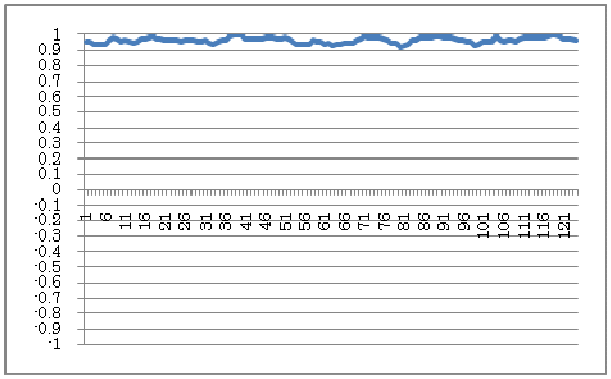

Fig. 12 experimental result for "Replacement" (HWT)

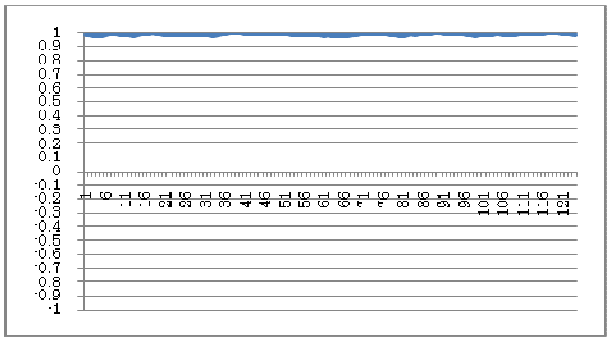

Fig. 13 experimental result for "Replacement" (HT)

C. *Experiment 3: ("Shuffle" ②)*

As in the "Replacement" experiment, no significant difference was observed in changing n, so we only show the result for n=5. In the graphs, the horizontal axis represents the number of shifts for the shuffled characters.

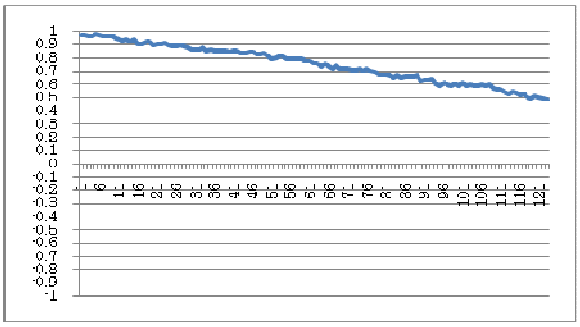

Fig. 14 experimental result for "Shuffle" (no transform)

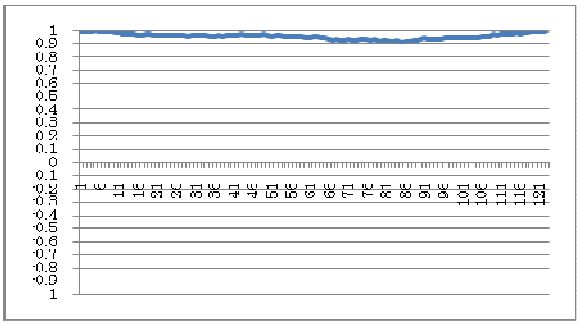

Fig. 15 experimental result for "Shuffle" (DFT)

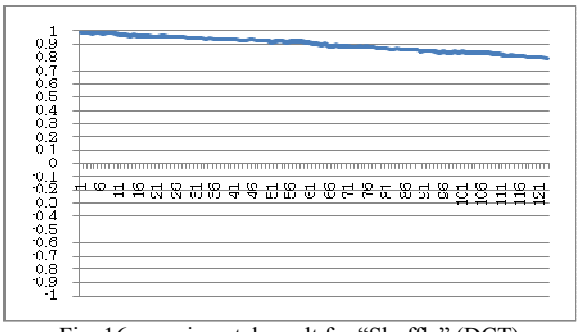

Fig. 16 experimental result for "Shuffle" (DCT)

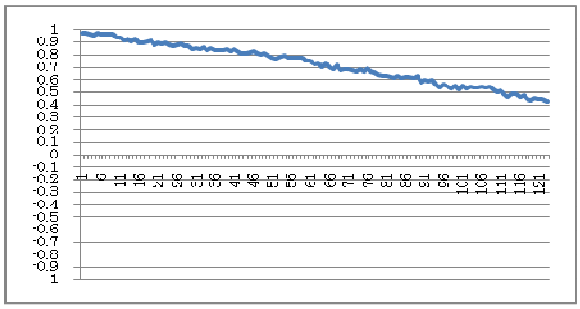

Fig. 17 experimental result for “Shuffle” (HWT)

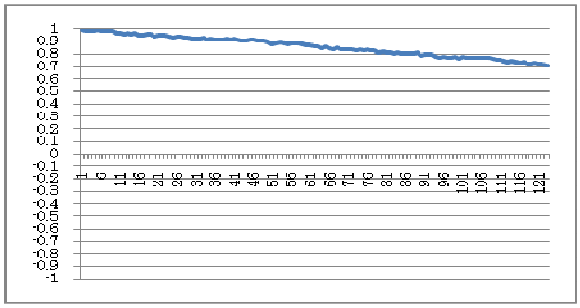

Fig. 18 experimental result for “Shuffle” (HT)

In Fig. 14 and 16 to 18, the larger the shifts be, the lower of the correlation coefficients become. The correlation reduces to 0.48 for no transformation, to 0.79 for DCT, to 0.43 for HWT and to 0.71 for HT. These values are almost the same as the average values for double-byte characters shown in Tables 1 to 5 for each of the transformations, respectively. For just n=5, a delimited number of characters reach the end of the sentence from the beginning. In the case that the sentence is arbitrarily divided into two, (e.g. when ‘あいうえ’ becomes ‘うえあい’) the correlation coefficients are almost identical to those of dissimilar documents.

In Fig . 15 (DFT), the decreasing trend stops and correlation coefficients begin to increase when the “time”-lags are more than 80. In all the cases the correlation coefficients are eventually asymptotic to 1. For various n, when a document that is divided into is shuffled, the correlation coefficients always become 1.00.

The lowest correlation coefficient of 0.92 exceeds the average value for double-byte characters shown in Table 2.

D. Experiment 4: (“Insertion” ③)

The results for the "Insertion" experiment are shown below. The similarity of the document is measured by the autocorrelation coefficient here as well. In this experiment, each time shifting is performed; the number of data points is reduced.

For the HWT and HT, correlation can be taken only when the number of data points is a power of two; hence the result is as shown Fig. 22 and 23. In all the figures, the horizontal axis represents the “time”-lags in the input data.

Focusing on the untransformed, DFT and DCT, the results are shown in Fig. 19 to 21. It can be observed that, except for the DFT, the correlation coefficients decrease rapidly. As for the DFT, the decrease is small. In this experiment, when the lag is 52, the correlation coefficient is 0.88, which is below the average for double-byte characters shown in Table 2.

In Experiment 4, it can be observed that as the “time”-lag increases, the magnitude of the correlations increases. This is probably because the data points reduction made the sensitivity high so that a slight change in the coded character results in a large change in the magnitudes.

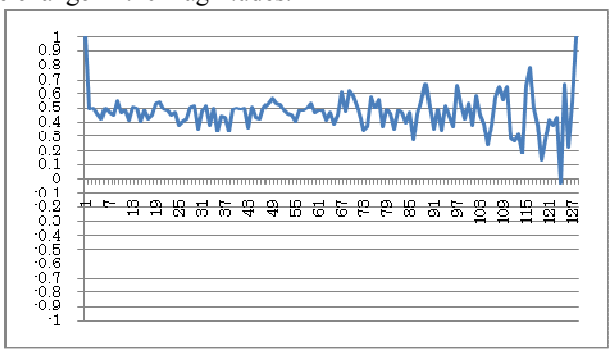

Fig. 19 experimental result for “Insertion” (no transform)

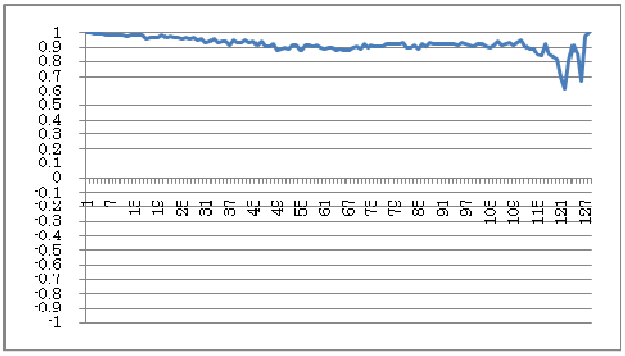

Fig. 20 experimental result for “Insertion” (DFT)

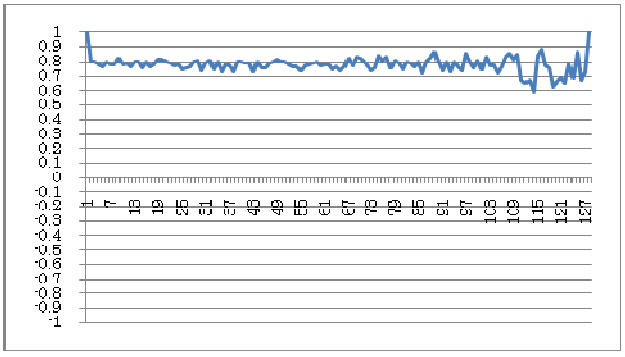

Fig. 21 Experimental result for “Insertion” (DCT)

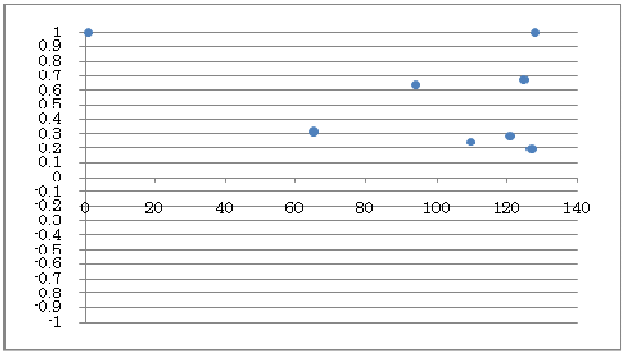

Fig. 22 experimental result for “Insertion” (HWT)

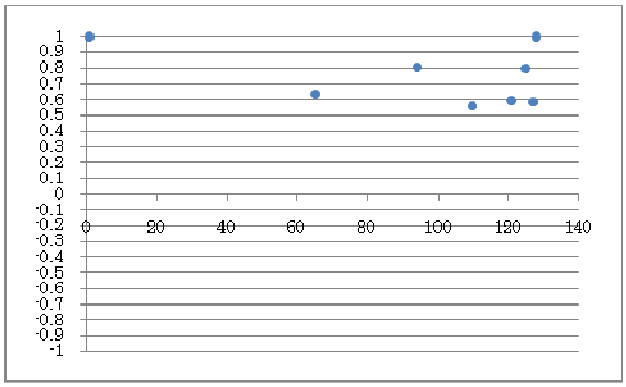

Fig. 23 experimental result for "Insertion" (HT)

VI. Experiment with a Real Plagiarized Documents

In this section, we evaluate the performance of the proposed method against real-world plagiarized documents. The input data is the original and plagiarized documents introduced in Section IV-A and this time around experiments were performed on 3 sets of data that included the example documents.

Input Data 1 is double-byte characters only （Japanese only) document, and the plagiarism feature that can be seen 'Replacement' ①. Input Data 2, consists of mixed single-byte and double-byte characters (alphabet), and the plagiarism feature that can be seen 'Replacement' ①. Additionally, because of the shift when the alphabet portion was transformed from double-byte to single-byte characters, the data also possesses the 'Insertion" ③ feature. Input Data 3 consists of mixed single-byte and double-byte characters (Arabic numbers), and the plagiarism feature that can be seen 'Shuffle' ②. The experimental results are shown in Table 6.

.

TABLE XI. correlation coefficients of the various types transform for plagiarized document

	NT	DFT	DCT	HWT	HT
Data1	0.818	0.961	0.941	0.801	0.940
Data2	−0.224	0.929	0.336	−0.148	0.373
Data3	−0.273	0.859	0.510	0.087	0.588

Input Data 1 in Table 6 is comparison with double-byte characters only, and when similarity comparisons are made, the corresponding results are those of "Double-byte" portion in Tables 1 to 5. Similarly, Input Data 2 and 3 correspond to the "Mixed-byte" portion in Tables 1 to 5.

Result of Input Data 1 show high values when compared to Experiment 1, irrespective of whether a transform is applied or not. This result is similar to that of the "Replacement" in Experiment 2 where the change is small, and it can be considered that the documents of double-byte characters only exhibit high correlations. For the Input Data 2, when the Fourier transform is compared with Experiment 1, an even higher variation in STD is obtained. For Input Data 3, the correlation coefficients are higher than those for the Fourier and Hadamard transforms in Experiment 1. For the Hadamard transform, there is a range of values where dissimilar documents are applicable, and for the Fourier transform, the variation in the range exhibits higher values. In the experiment, the effectiveness of the Fourier transform is further demonstrated.

VII. Character-code-conversion

In the experiment of real document, we had a problem. The document is high similarity but correlation became low in case the comparison to the text in which the single-byte character and the double-byte character are intermingled.

Moreover, in the document of only double-byte character, even if reading was the same, the problem which cannot perform judgment of similarity correctly from the difference in expression of Hiragana, Katakana and Chinese character. As the reason, it is because character codes differ by the difference between single-byte and double-byte, and the difference in Hiragana or Katakana.

The example of the character in which character codes differ by same reading is shown in Fig. 24.

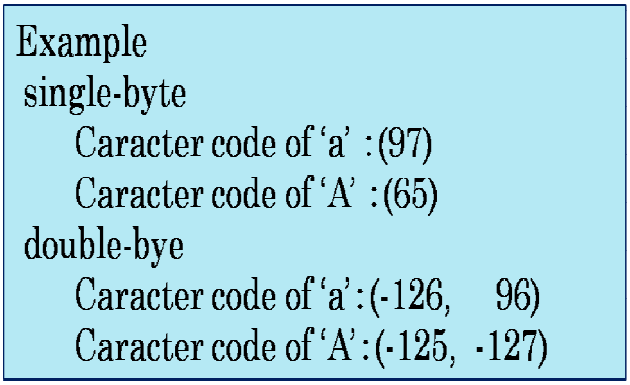

Fig. 24 difference of character code by difference of expression

It turns out the character codes difference when the kinds of character codes differ by fig.24.The document is various according to person and the scene to write. Selection whether to use Hiragana or Katakana differ the cases where it is used. Especially for Japan, since we use various kinds of characters about Hiragana, Katakana, Kanji, and the alphabet there are some methods expressing the same words. In order to raise the accuracy of similarity judgment, we propose a new method is character-code-conversion. The character-code-conversion unifies into one character code the character code of the Hiragana, the Katakana and the alphabet which is the same reading. In a Kanji, in order that many may carry out two or more reading, character-code-conversion is not adapted. Correspondence of character-code-conversion is shown in Fig. 25.

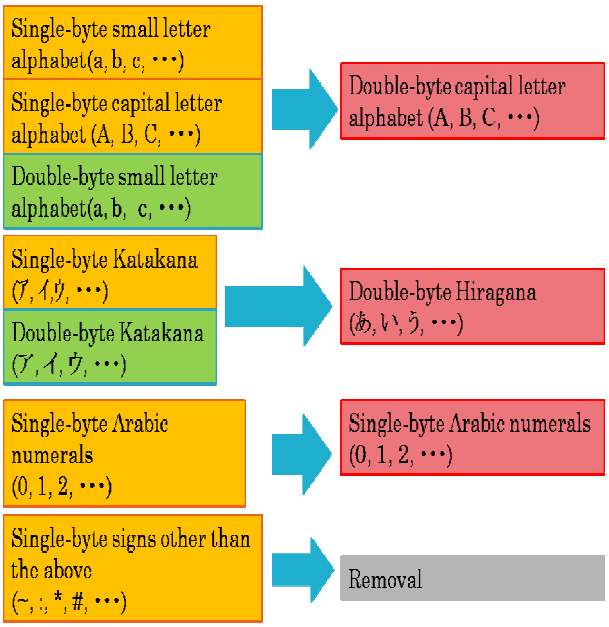

Fig. 25 character-code-conversion

The alphabet has a single-byte small letter, a double-byte capital letter, a double-byte small letter, and a double-byte letter, and it can express it by four kinds of character codes to each character. Katakana and a Hiragana have single-byte Katakana, double-byte Katakana, and a double-byte Hiragana, and it can express them by three kinds of character codes to each character. Arabic numerals have single-byte Arabic numerals and double-byte Arabic numerals, and it can express them by two kinds of character codes to each character.

In the method proposed this time, we convert the character as follows: the alphabets are unified into the double-byte capital letter alphabets, Hiragana and Katakana are unified into a double-byte Hiragana, and Arabic numerals are unified into double-byte Arabic numerals as shown in Fig. 25. We convert all into the double-byte character because the character which can be denoted by a single-byte character can be expressed by a double-byte character. By unifying all into a double-byte characters, we can improve the problem over mixture of the single-byte characters and the double characters. By comparing document by the character code of the same domain, improvement in the accuracy of a similarity judgment is expectable.

The single-byte signs other than single-byte alphabet, single-byte Katakana, and single-byte Arabic numerals are removed when changing. As the first reason, it is because the half-width sign and the full-size sign do not necessarily one-to-one correspondence. As the second reason, the rate of a sign used is low and it is seldom subject to the influence of the character-code-conversion.

Based on the above-mentioned conditions, we conducted the experiment which investigates the shift of the correlation coefficient by character-code-conversion. We conducted two kinds of experiments that are the case of conversion into double-byte from double-byte and the case of conversion double-byte width from single-byte. The comparison method is shown in Fig. 3.

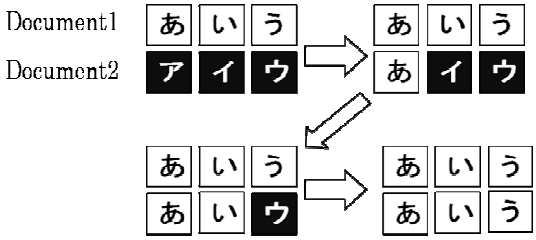

Fig. 26 the comparison method

Fig. 26 shows the example of the comparison to the text of the double-byte of three characters. One side is a text after conversion and another side is a document before conversion. In the process changing the document of one character before conversion at a time as shown in Fig. 26 we investigate how the correlation coefficients shift.

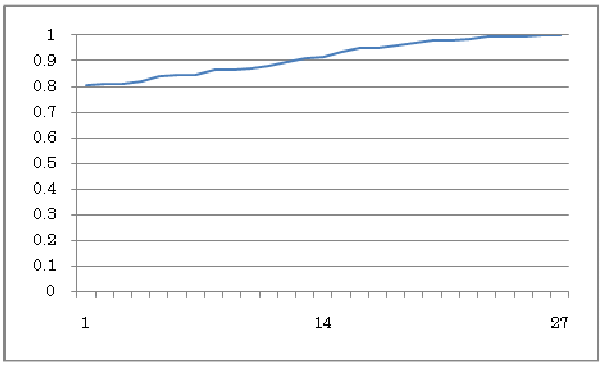

Fig. 27 shift of the correlation coefficients to double-byte from double-byte

Fig. 27 shows the experimental result of the double-byte of 26 characters. This graph shows that the correlation coefficients increase at a constant rate.

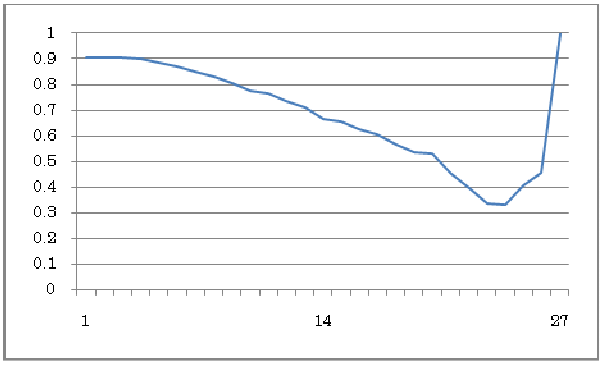

Fig. 28 shift of the correlation coefficients to double-byte from single-byte

The graph of Fig. 28 shows shift of the correlation coefficients at the time of converting the single-byte small letter alphabet into the double-byte capital letter alphabet. The compared number of characters is 26. The correlation coefficients decrease from about 0.9, and it takes the lowest value when 24 characters are converted. A correlation coefficient increases by

a rate of change higher than the time of decreasing after that. And finally it increases to 1.

From this result, character code conversion shows that the accuracy of the similarity judging between documents improves. Therefore, character code conversion is a method effective in the similarity judging between documents.

VIII. Change of the Correlation Coefficient by the Length of a Sentence

The number of characters of a text varies with many factors. Obviously, it can be say the same for the sentence which constitutes a text. For example, the length of a sentence changes with the uses from the comparatively short sentence seen by e-mail to the long sentence seen by the novel. It may over 300 characters, if it is a long sentence in Japan.

In this research, in order to make the similarity of a document judge, we define a correlation coefficient as the degree of similar. In this section, we define the threshold in order to judge whether similarity is high or low. However, since the variations in the correlation coefficients differ with the number of characters, i.e., the number of data, a threshold cannot be decided in one. The experiment which investigates change of the correlation coefficients by the length of the number of characters to compare is conducted first, and then the thresholds of the judgment of similarity are defined for every length of a sentence.

We explain the experiment. In this experiment, we use only Fourier transform. Two sentences of the same number of characters are prepared. Changing the number of characters of the sentence from 2 characters to 500 characters, we investigate change of the correlation coefficients. The sentence to compare is a text whose similarity is very low because of that the characters chosen from 6891 kinds of characters at random. The characters to be used are the double-byte characters after the character-code-conversion containing the Hiragana, the Kanji, and the alphabet. 100 trial is performed to each number of characters of 2 to 500 characters respectively. In each number of characters, Fig. 29 shows the average value of the 100 correlation coefficients, and Fig. 30 shows the standard deviation of the 100 correlation coefficients. The horizontal axis of the graph shows the comparing number of characters.

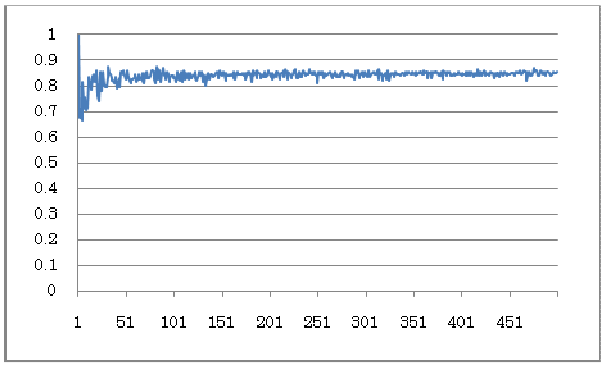

Fig. 29 average of the correlation coefficients with change of the number of characters

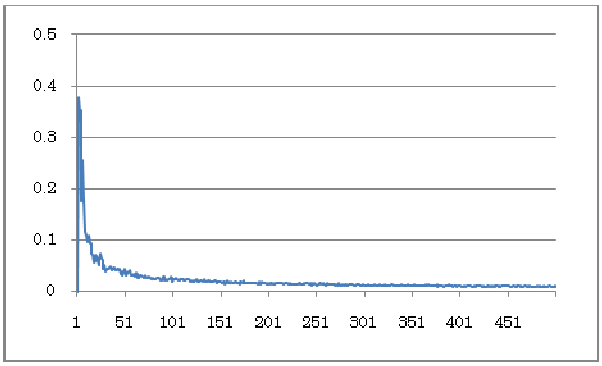

Fig. 30 standard deviation of the correlation coefficients with change of the number of characters

In Fig. 29, variation is seen by change of average value when there are few characters to compare. In response, the standard deviations shown in Fig. 30 indicate a big value when the number of characters is low.

In Fig. 29, the average of the graph comes out 0.840. However, since variation becomes large when the number of characters is low, it is necessary to decide a threshold for every number of characters.

The determination of a threshold to each number of characters is described. The maximums of the correlation coefficients which it obtained each by 100 comparisons are plotted. The especially high values in it are connected by a straight line, and the straight line is determined as a threshold value. However, a threshold value is arbitrarily defined as an exception what has a few number of characters to compare. In one to six characters, the thresholds are determined as 1 because the maximum turns into a value very near 1. In seven to seventeen characters, the thresholds are determined as 0.965 similarly. Fig. 8 shows the graph of the threshold.

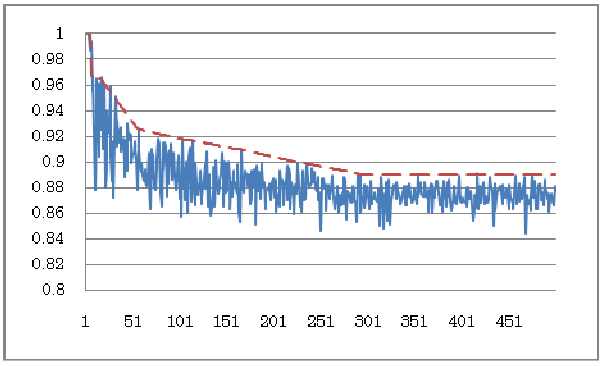

Fig. 31 threshold for the judgment of document similarities

In Fig. 31, what was shown with the dashed line is the threshold obtained in this experiment. Decrease ratio of the threshold is the highest among 17 to 56 characters. Then, the threshold decrease gently among 56 to 290 characters, and the line become fixed. Decrease ratio of the threshold is the highest among 17 to 56 characters. Then, the threshold decreases gently among 56 to 290 characters, and the line become a constant value. This threshold is the line created from the highest correlation coefficients when a sentence with very low similarity is compared 100 times. Therefore, in a domain higher

than this line, it is judged that it is a high similar document. If the number of characters exceeds 300, change of standard deviation will become almost fixed. Thus, in the comparison to the document of 300 or more characters, even if it defines a threshold value uniformly, it can be said that it is satisfactory. Additionally the text of over 300 characters is rare if it is assumed to comparison to 1 text, and the evaluation to the text of over 300 characters is enough if it refers to the value near 300 characters.

The threshold value obtained in this experiment is a value which referred to the high value of especially correlation coefficients and defined it when a text with low similarity is compared. Therefore, if a comparison result becomes higher than a threshold value, it can be expected that document similarity is a high.

IX. Summarize

Putting everything together, the experiments can be summarized as follows.

In a section V-A, using dissimilar documents, the values that are used as a standard to measure the performance of the proposed method are obtained.

In a section V-B, it is shown that when the change in the number of characters is small, high values are obtained independently of the type of orthogonal transform. However, this result is under the condition that the number of characters is 128, the power of two.

In a section V-C, for the "Shuffle" experiment, it is shown that the Fourier Transform is effective for measuring similarity. The reason that the DFT values always become 1.00 when the document is divided into two could be that the Fourier transforms works on the assumption that the input signal is periodic. When the input signal points are separated arbitrarily, a phase change occurs but there is no change in the amplitude. For that reason, the frequency of the original document becomes equal to that of the modified document.

In a section V-D, the Fourier Transform is shown to be effective for shifted document. For this reason, as in Section V-C, it can be considered that the Fourier transforms treats the input signal as a periodic function.

In a section A, we propose a method that is character-code-conversion. Character-code-conversion is especially effective the document that is used mixed-byte characters.

In a section B, we determine a threshold line for judgment of document by comparison of dissimilar document.

X. Conclusion

In this paper, we proposed one simple method of measuring the similarity of Japanese documents without resorting to morphological analysis, but instead we performed orthogonal transform on the data sequence in order to get the correlation characteristics, and then went on to thoroughly investigate the characteristics of the correlation coefficients. The results can be summarized as follows.

For the document on which insertion was performed, different characters are plugged into the document and the entire text is shifted. In this case, the Fourier transform was found to be effective. Additionally, this effectiveness was confirmed with real plagiarized documents.

In experiment of real document including mixed-byte characters, suitable judgment was not conducted. We proposed a new method that is character-code-conversion and experimented using it in order to solve that problem. By character-code-conversion introduction, improvement in accuracy is expectable to the document containing mixed-byte character.

Furthermore, we determine a threshold in order to judge the document similarity. In this method, we use correlation coefficient as the degree of similar. Thus, the threshold corresponding to the number of characters will make a more outstanding judgment.

As a future work, it is assumed to examine various comparison methods. And we will conduct experiment using those methods. Moreover, we will deal with various real documents not only Japanese as input data, and examine the property.

Acknowledgment

A part of this work was supported by MEXT Grant-in-Aid for Building Strategic Research Infrastructures. We are very grateful for their support.

References

[1] T. Chikayama, K. Yanauchi, M. Miwa, “Attempt of Judgment Similarities for Document Without knowledge of Natural Language”, programming-symposium report collection of information processing society in summer, pp.81-86, 2003 [in Japanese].
[2] N. Wada, “Well Understanding Signal Process”, pp.79-84, Morikita publisher 2009.
[3] Y. Kogure, “Understanding Fourier System”, Koudansya Co. 1999.
[4] S. Kiya, “Digital Signal Process”, Shoukoudou Co., pp.95-99, 1997.
[5] Version Blog of Mandana Communication, http://mandanatsusin.cocolog-nifty.com/blog/2007/06/post_a49c.html

In this study, experiments were mainly performed on double-byte characters. As future research, we would carry out experiments with single-byte characters.