§35. Study on Quick Similarity Search in Massive-size Waveform Databases

Hochin, T. (Kyoto Institute of Technology),
Nakanishi, H., Kojima, M.

Experiments of the fusion phenomena produce a lot of sequences of time-varying values which form waveforms. If the waveforms similar to a desired one can be obtained by using computer system, the burden of researchers in searching similar waveforms will be extremely decreased. We have addressed to the issue on this kind of retrieval. The Fourier series coefficients of waveforms are used in calculating the dissimilarity of waveforms. For the slowly-varying waveforms, several low-frequency coefficients are enough to decide the similarity of waveforms[1]. The Euclid distance of the power spectrum density is efficient in retrieving the waveforms having time-sectional oscillation patterns[2]. These methods have successfully improved the performance in retrieving similar waveforms. These methods, however, could not correctly retrieve similar severely-varying waveforms. An example of this type of waveform is shown in Fig. 1. As shown in Fig. 1, high-frequency components are dominant in this type of waveform. This may result in the deficiency of the dissimilarity (distance) of waveforms.

This paper treats severely-varying waveforms, and proposes two extensions to the dissimilarity of waveforms in order to improve its correctness in retrieving severely-varying waveforms. The first extension is to capture the difference of the importance of the Fourier series coefficients of waveforms against frequency. The Fourier series coefficients are divided into three groups: the low frequency group, the middle frequency one, and the high one. Treatment of coefficients differs according to the frequency group that the coefficients belong to. For the coefficients in the low frequency group, the Euclid distance of the coefficients themselves is contributed to the dissimilarity. For those in the middle one, the Euclid distance of the power spectrum density[2] is contributed to the dissimilarity. For those in the high one, the difference between the average values of the power spectrum density is contributed to the dissimilarity. The second extension is to consider the outlines of waveforms. The dissimilarity includes the Euclid distance of the Fourier series coefficients of the outline of a waveform. The dissimilarity, which is called the *spectrum distance considering frequency and outlines*, is defined as follows:

$$D_f^{ao}(x,y) = r_1\left(\sqrt{\sum_{f=1}^{k-1}\left|X_f - Y_f\right|^2} + \sqrt{\sum_{f=k}^{l-1}\left(\left|X_f\right| - \left|Y_f\right|\right)^2}\right) + r_2\left(\sqrt{\sum_{f=1}^{k-1}\left|X''_f - Y''_f\right|^2} + \sqrt{\sum_{f=k}^{l-1}\left(\left|X''_f\right| - \left|Y''_f\right|\right)^2}\right) + r_3\sqrt{\sum_{j=1}^{d}\left(\sum_{f=i_{j-1}+1}^{i_j}\left(\left|X''_f\right| - \left|Y''_f\right|\right)\right)^2}$$

Here, X''_f is a coefficient of the Fourier series of an outline of a waveform $[x_t]$.

The correctness of the proposed dissimilarity is experimentally evaluated. Waveforms similar to the waveform shown in Fig.1 are retrieved from 1000 waveforms. These are the magnetic waveforms obtained through the experiments at National Institute for Fusion Science. Four kinds of distances are compared: the Euclid distance, the spectrum distance, the gradually averaged spectrum distance, and the proposed one. The third one is the special case of the proposed distance that does not consider outlines, i.e., $r_1=r_3=1$, $r_2=0$, $X''_f=X_f$, and $Y''_f=Y_f$. The metrics used in the evaluation are precision and recall. Precision (recall, respectively) is calculated by N_{sim_ret}/N_{ret} (N_{sim_ret}/N_{sim}) where N_{sim_ret} is the number of the similar waveforms retrieved, N_{ret} is the number of the retrieved waveforms, and N_{sim} is the number of the similar waveforms. In the retrieval based on the gradually averaged spectrum distance, the parameters and their values are k=1, l=49, and d=4, while those in the retrieval based on the spectrum considering frequency and outlines are k=7, l=57, d=8, $r_1=1$, $r_2=5$, and $r_3=6$.

The precision and recall curves obtained are shown in Fig. 2. The retrieval based on the proposed distance outperforms those based on the other kinds of distance. The retrieval based on the spectrum distance outperforms the retrieval based on the Euclid distance. The difference that we can not identify with our eyes may influence the distance. Considering the outline brings us the good performance. As an outline varies slowly, we could easily capture the shape of the outline.

Reference

1) H. Nakanishi, H., Hochin, T., Kojima, M., LABCOM Group : Search and retrieval method of similar plasma waveforms, Fusion Eng. and Design, 71 (2004) 189-193.
2) H. Nakanishi, T. Hochin, M. Kojima, LABCOM Group: Similar pattern search for time-sectional oscillation in huge plasma waveform database, Fusion Eng. Design, 81 (2006) 2003-2007.

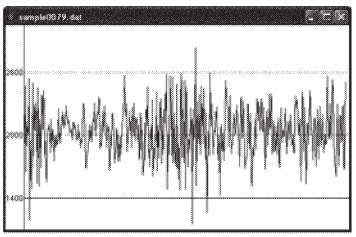

Fig. 1. Key waveform.

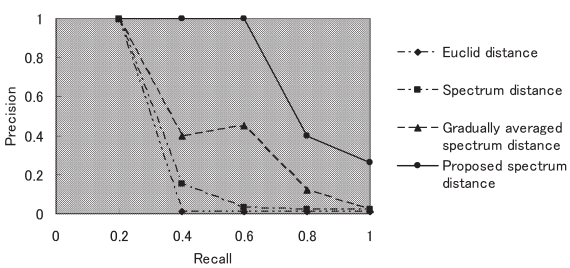

Fig. 2. Result of performance evaluation.