§39. 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 extremely be decreased. We have addressed to the issue on this kind of retrieval. Fourier series coefficients of waveforms are used in calculating the dissimilarity of waveforms. In order to make the subsequence matching of waveforms possible, a waveform is divided into segments. Fourier series coefficients are obtained for each segment. The multi-dimensional index is constructed by using these coefficients. Fourier series coefficients are also stored into a file, which is called a seek file, according to the order of the segments[1]. By using this seek file, the time of consulting the index can be decreased. This could make the retrieval performance better.

Moreover, we have proposed two extensions to the dissimilarity of waveforms in order to improve its correctness in retrieving severely-varying waveforms[2]. The first extension is to capture the difference of the importance of Fourier series coefficients of waveforms against frequency. The second extension is to consider the outlines of waveforms. The dissimilarity, which is called the *spectrum distance considering frequency and outlines (SDFO)*, 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 Fourier series of an original waveform $[x_t]$, X''_f is a coefficient of Fourier series of an outline of the waveform, and r_i is a weight coefficient of a term of the equation. The first term of the equation is on the distance of the low frequency band and that of the middle one of an original waveform. The second term is on the distance of the low frequency band and that of the middle one of the outline. The third one is on the distance of the high frequency band of the outline.

The storage amount, the construction time, and the retrieval one of the indexing method using a seek file and SDFO are experimentally evaluated. The magnetic waveforms obtained through the experiments at National Institute for Fusion Science are used in the experiments. The number of waveforms is varied up to 300. Each waveform is divided into 256 segments, each of which contains 512 points. In the retrieval based on SDFO, the parameters and their values are k=7, l=57, d=8, r_1=1, r_2=5, and r_3=6. The total number of the coefficients used is 133.

The size of the index file and that of the seek file are measured in the storage amount evaluation. The results are shown in Fig. 1. The size of index file is extremely larger than that of the seek file. This may be caused by the large number of the coefficients of a segment.

The construction time of the index file for 300 waveforms is about 2,830 seconds, while that of the seek file is about 1.2 second. Constructing an index file requires a lot of time.

Figure 2 shows the retrieval time in the sequential search and that in the index search. In the sequential search, the more data are stored, the more time is required. This may be caused by the sort of the result data according to the similarity. In this experiment, the number of dimensions is 133 because the number of the coefficients used is 133. Although this is bigger than 10, the index search overcomes the sequential search.

1) T. Hochin, H. Nakanishi, M. Kojima: On the Construction of Databases of Experiment Data, Ann. Rep. NIFS (2005-2006) 165.
2) T. Hochin, H. Nakanishi, M. Kojima: Study on Quick Similarity Search in Massive-size Waveform Databases, Ann. Rep. NIFS (2006-2007) 175.

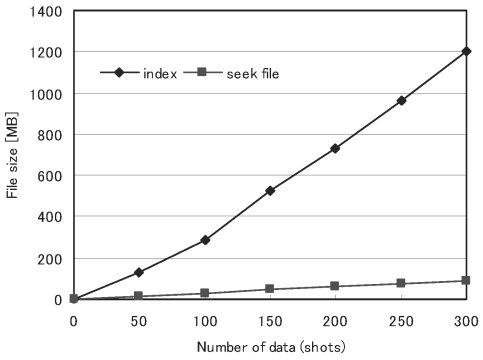

Fig. 1. Result of storage evaluation.

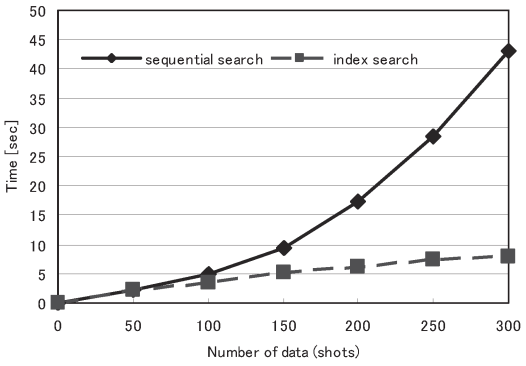

Fig. 2. Result of retrieval performance evaluation