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

Hochin, T., Nomiya, H. (Kyoto Inst. of Tech.),
Okumura, H. (Mie Univ.),
Nakanishi, H., Kojima, M., Nagayama, Y., Ohdachi, S., Emoto, M., Ohsuna, 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. [1] We have proposed an indexing method of the severely and quickly changing plasma waveforms for accelerating search and retrieval of their subsequences. [2] The method divides a waveform into fine-grained segments. The similar segments are grouped into a segment group, which is managed by using a multi-dimensional index. A sequence of segments is used as a unit in matching subsequences in the retrieval. This method could bring us the quick construction and the small size of the index, and the good performance and the good retrieval correctness. [2]

This paper treats movies of plasma discharge, which are called plasma movies. Although a plasma movie is a kind of time series data, it could not be treated well through the conventional methods of handling time series data because a unit of a movie is a picture (frame).

A plasma movie is divided into segments as in the method described above. [2] A segment includes sixteen frames. The duration of a segment is about 0.5 second. The frame is of the center region of the original frame for the purpose of excluding the character strings of shot number, and so on. The width and the height of a frame are 256 pixels and 128 pixels, respectively. Feature values described later are obtained from each segment. By using the feature values, dissimilarity of the key segment of a plasma movie and a segment of a plasma movie stored in a database is calculated. The absolute values of the Fourier coefficients obtained by applying three-dimensional Fourier transformation to the luminance values of a segment are used as the feature values. Ten coefficients are selected because many coefficients are obtained. We have tried three methods for this selection. The first method uses variances of the absolute values of the Fourier coefficients obtained from 82 segments out of five typical plasma movies. Ten coefficients, which have the largest variances, are selected. The second method uses the coefficient of variation, which is the standard deviation divided by the average value, of the absolute values of the Fourier coefficients. The last method filters the coefficients by using their variances, then select ten coefficients whose coefficients of variation are the largest.

Correctness of retrieval is evaluated by using ten plasma movies (totally 120 segments). A snapshot of the key segment, which is the third segment of the plasma movie whose shot number is 40215, is shown in Fig. 1. It is evaluated with recall (= n_{cor} / n_{sel}) and precision (= n_{cor} / n_{corall}), where n_{cor} is the number of the correct segments retrieved, n_{sel} is the number of the segments retrieved, and n_{corall} is the total number of the correct segments. The correct segments are manually decided from 120 segments. The number of the correct segments is 14.

The recall-precision curves are shown in Fig. 2. In Fig. 2, the method using the Fourier coefficients selected by using the variance (coefficients of variation, respectively) is depicted by “Variance” (“COV”). The method using the Fourier coefficients selected by using the coefficients of variation after the filtering with variances is depicted by “Two steps.” In the “COV” method, the first two candidates are correct, while the followings include errors. The “variance” method retrieves many errors in the first several candidates. The first several candidates are not correct in the “two steps” method, while the latter ones are correct. This brings the good accuracy to this method.

The retrieval key segment varies from dark to bright, while the incorrect segments retrieved as the first several candidates vary from bright to dark. This may be caused by the fact that the phase difference is not considered in the feature values, which are the absolute values of the Fourier coefficients.

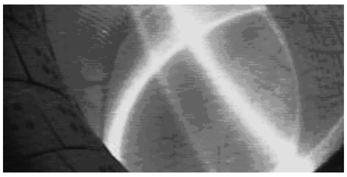

Fig. 1. A snapshot of the key segment.

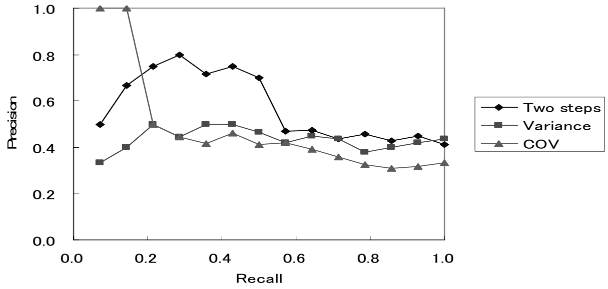

Fig. 2. Evaluation result of retrieval accuracy.

1) Hochin, T., Koyama, K., Nakanishi, H., Kojima, M., LABCOM group: Extension of frequency-based dissimilarity for retrieving similar plasma waveforms, Fusion Eng. Des. **83** (2008) 417-420

2) Yamauchi, Y., Hochin, T., Nomiya, H., Nakanishi, H., Kojima, M.: Fast Partial Similarity Retrieval Method of Swinging Time Series, IPSJ SIGDBS **2009-DBS-148** (17) (2009) 1-8 (in Japanese)