§33. 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. The multi-dimensional index is constructed by using these coefficients. Moreover, we have proposed the dissimilarity of waveforms called the spectrum distance considering frequency and outlines (SDFO) in order to improve correctness in retrieving severely-varying waveforms[1].

This paper proposes a method of subsequence matching of plasma waveforms. The proposed method divides a waveform into fine-grained segments having the same length w. A segment is a series of points. A waveform is treated as a series of segments. The segment length w is shorter than the minimum query subsequence length l, and is equal to or nearly equal to the length u of the window movement of the sliding window approach. The proposed method uses the SDFO for the dissimilarity of subsequences for good retrieval accuracy. As the consecutive segments often have similar feature values, these segments are treated as a *segment group*. Segment groups are stored into an index. This index is called the *Segment Group (SG) index*. In retrieving subsequences, a sequence of several segments, which is called a *section*, is used as a unit in matching subsequences. The average values of the segments in a section of a query subsequence are used in obtaining the segments from the SG index. From the segments obtained, the consecutive segments are selected to form a subsequence. The dissimilarity of the query subsequence and a subsequence is the sum of the distances of the corresponding sections. The average values of the segments in a section are used in calculating the distance. For overcoming the shift errors of subsequences, sections are overlapped each other.

The proposed method is evaluated by comparing it with the ST index and the prefix search method[2]. In the comparative method, feature values are obtained from the points in a window slided forward by an offset, and are grouped and stored into a multi-dimensional index. The first u points in a subsequence are used in the retrieval. In this evaluation, the SDFO is used as the dissimilarity measure of the comparative method. The waveforms used in the experiments are of the magnetic field fluctuations obtained through the experiments at National Institute for Fusion Science. A waveform (segment, respectively) is consisted of 131072 (512) points. A waveform is divided into 256 segments. The minimum query subsequence length l is 4096. The length and the offset of a section are 8192 and 4096, respectively.

The times in retrieving subsequences, whose lengths are 32, 64, 96, 128, 160, and 192 segments, are measured. The times in retrieving subsequences are shown in Fig. 1. For the short subsequences, the retrieval times are almost same in both methods. The longer the subsequence is, the more time is needed in the comparative method.

Correctness of the retrieval is evaluated by using the metrics of the information retrieval, i.e., precision and recall[3]. The precision-recall curves are shown in Fig. 2. The proposed method keeps precision high until 0.7 of recall, while it becomes drastically low around 0.2 of recall in the comparative method. The correctness of the proposed method is better than that of the comparative one. This may be caused by that the unit of comparison is a section, and sections are overlapped.

1) T. Hochin, K. Koyama, H. Nakanishi, M. Kojima, and LABCOM group, Extension of frequency-based dissimilarity for retrieving similar plasma waveforms, *Fusion Eng. Des.* (2008) **83**(2-3):417-420

2) C. Faloutsos, M. Ranganathan, and Y. Manolopoulos, Fast subsequence matching in time-series databases, *Proc. of ACM SIGMOD 1994* (1994) 419-429

3) M. Kobayashi and K. Takeda, Information retrieval on the web, *ACM Comp. Surv.* (2000) **32**(2):144-173

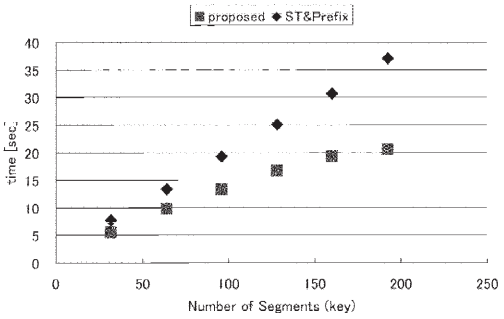

Fig. 1: Times of retrieving subsequences.

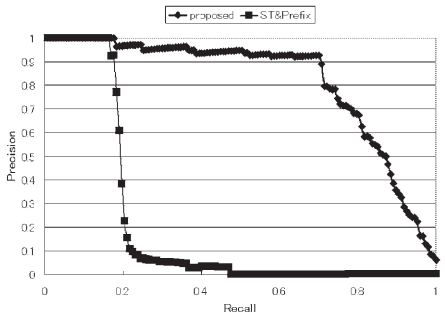

Fig. 2: Precision-recall curves.