§33. Improvement of Similarity Retrieval in Fusion Experiment Multimedia Data Archives

Hochin, T., Nomiya, H. (Kyoto Inst. of Tech.), 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 proposed an indexing method of the severely and quickly changing plasma waveforms for accelerating search and retrieval of their subsequences [1]. It is, however, not possible to obtain the waveforms similar to those stretched.

Rafiei *et al.* have proposed the method of retrieving waveforms transformed by scaling and translation [2]. Waveforms are retrieved by using a key waveform q and a set of transformations T. The results are obtained as a set of pairs of the form (w, t), where w is a waveform and t is a transformation in T. They have also proposed the method that a set of transformations is divided into several clusters, and the clusters are managed by using an index called the MT index. Although the retrieval time drastically changes according to the number of clusters, the optimal number of clusters has not been clarified. This study proposes the method of determining the optimal number of clusters by introducing the model of retrieval time.

The retrieval time t_{query} is the sum of the index search time t_{search} and the post-processing time $t_{postproc}$. In using the MT index, search is repeated N_C times, where N_C is the number of clusters. Given an index search time $t_{search\_on\_index}$, t_{search} is obtained by $N_C\ t_{search\_on\_index}$. As $t_{postproc}$ is proportional to the number of candidates N_{cand}, $t_{postproc}$ is obtained by $N_{cand} t_{conf\_ans}$, where t_{conf_ans} is a confirmation time. N_{cand} is proportional to $N_D\ N_T\ P_h$, where N_D is the number of data, N_T is the number of transformations, and P_h is the hit probability. P_h is proportional to the ratio of the sum of the volumes of the hyper-rectangles V_R to the volume of the whole space. Assume that the width of each dimension of the hyper-rectangle V_R is determined by the number of transformations a cluster. V_R is proportional to $(N_T/N_C)^n$. There are $N_D N_C$ hyper-rectangles. The sum of the volumes of hyper-rectangles is proportional to $(N_T/N_C)^n N_D N_C$. The number of candidates N_{cand} is proportional to N_C^{1-n}. Therefore, t_{query} is obtained by $N_C\ t_{search\_on\_index} + c\ N_C^{1-n}\ t_{conf\_ans}$, where c is a constant. The minimum value of N_C is obtained by $(c(n-1)\ t_{conf\_ans} / t_{search\_on\_index})^{1/n}$.

In order to confirm the validity of the model of the retrieval time described above, an experiment is conducted. The data series are obtained by adding the random value ranging from -500 to 500 one by one. The length of a series is 256. The number of series is 6000. A query series is randomly selected every time. The sets of transformations used are those of moving averages of 1 to 50, 1 to 75, and 1 to 100 points. The feature values of a series are the first and the second Fourier coefficients. The number of dimensions is four. The numbers of clusters are the divisors of the number of transformations. The average value of the times of 1000 retrievals is obtained.

The retrieval time t_{query} is shown in Fig. 1. The retrieval time t_{query} is concave. When N_c is equal to 5, the retrieval time is the smallest for all of transformation sets. In the experimental results, the index search time t_{search} is proportional to the number of candidates N_{cand}. N_{cand} is proportional to the power of -2.7 of N_c as shown in Fig. 2. The index search time $t_{search\_on\_index}$, the confirmation time t_{conf_ans}, and the constant c obtained from the experiment are shown in Table 1. The optimal numbers of clusters calculated are 4.2, 5.3, and 7.2 for the number of transformations N_T = 50, 75, and 100, respectively. As the number of clusters is a divisor of the number of transformations, the optimal number of clusters is 5 for all of the transformation sets. This meets the results of the number of clusters N_c measured as shown in Fig. 1.

1) Hochin, T., Yamauchi, Y., Nakanishi, H., et al.: Indexing of plasma waveforms for accelerating search and retrieval of their subsequences, Fus. Eng. Des., **85**(5) (2010) 649-654.
2) Rafiei, D. and Mendelzon, A.O.: Querying time series data based on similarity, IEEE Tr. on Know. and Data Eng., **12**(5) (2000) 675-693.

Table 1 Times and values obtained

N_T	$t_{search\_on\_index}$ [ms]	t_{conf_ans} [ms]	c
50	0.343	0.061	570
75	0.287	0.035	2194
100	0.289	0.040	6440

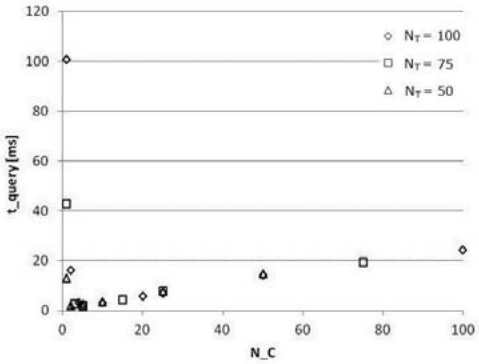

Fig. 1. Retrieval time t_{query}.

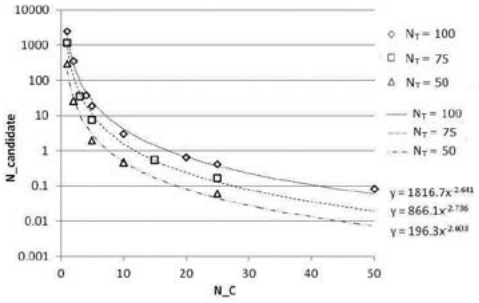

Fig. 2 Number of candidates N_{cand}