View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by National Institute for Fusion Science (NIFS-Repository)

§23. Study on Quick Similarity Retrieval in Massive-size Diagnostic 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. We have proposed an indexing method of the severely and quickly changing plasma waveforms for accelerating search and retrieval of their subsequences. Moreover, the method using three dimentional Fourier transformation has been developed for the similarity retrieval of movies of plasma discharges, which are called plasma movies.[1] This paper proposes the dissimilarity of plasma movies for improving retrieval accuracy.

We could distinguish the differences of phases at the low frequency band, while we could not distinguish them at the middle frequency band. The differences of frequencies could be distinguished at the middle frequency band, while those could not be distiguished at the high frequency band. In the case of high frequency, we could distinguish the difference of frequencies when the frequency is extreamly different from each other. The dissimilarity D considering this tendency is represented with the following formula:

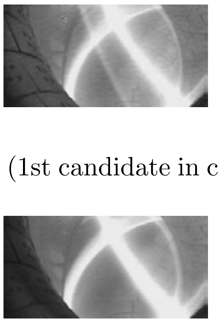
$$
\begin{aligned}
D &= C_{DC}|X_{DC} - Y_{DC}| \\
&+ C_{l0}\sqrt{\sum_{f_y=0}^{fyl(0)} \sum_{f_x=lowbeg(0,f_y)}^{low(0,f_y)} |X-Y|^2} \\
&+ C_{lN}\sqrt{\sum_{f_t=1}^{ftl} \sum_{f_y=0}^{fyl(f_t)} \sum_{f_x=lowbeg(f_t,f_y)}^{low(f_t,f_y)} |X-Y|^2} \\
&+ C_{m0}\sqrt{\sum_{f_y=0}^{fym(0)} \sum_{f_x=low(0,f_y)+1}^{mid(0,f_y)} (|X|-|Y|)^2} \\
&+ C_{mN}\sqrt{\sum_{f_t=1}^{ftm} \sum_{f_y=0}^{fym(f_t)} \sum_{f_x=low(f_t,f_y)+1}^{mid(f_t,f_y)} (|X|-|Y|)^2} \\
&+ C_{h0}\sqrt{\left(\sum_{f_y=0}^{fyh(0)} \sum_{f_x=low(0,f_y)+1}^{fxh} |X|-|Y|\right)^2} \\
&+ C_{hN}\sqrt{\left(\sum_{f_t=1}^{fth} \sum_{f_y=0}^{fyh(f_t)} \sum_{f_x=low(f_t,f_y)+1}^{fxh} |X|-|Y|\right)^2}.
\end{aligned}
$$

Here, the three dimentional frequency of a retrieval key movie (a retrieval result movie, respectively) is represented with $X = X(f_x, f_y, f_t)$ $(Y = Y(f_x, f_y, f_t))$. C_{l0} and C_{lN} are the coefficients at $f_t = 0$ and $f_t \geq 1$, respectively. The function $lowbeg(f_t, f_y)$ $(low(f_t, f_y)$, respectively) returns the minimum (maximum) value of f_x of the low frequency band at (f_t, f_y). The function $fyl(f_t)$ returns the maximum value of f_y of the low frequency band at f_t, while the function ftl returns the maximum value of f_t of the low frequency band. The others are similar to these ones.

The boundaries of the frequency bands are experimentally decided. The low frequency band is represented with $R_{L0} \cup R_{L1}$, where $R_{L0} = \{(f_x, f_y, 0) \mid 1 \leq f_x + f_y \leq 4\}$, and $R_{L1} = \{(f_x, f_y, 1) \mid 0 \leq f_x \leq 1 \wedge 0 \leq f_y \leq 1\}$. The middle one is represented with $R_{M0} \cup R_{M1} \cup R_{MM}$, where $R_{DC} = \{(0,0,0)\}$, $R_{M0} = \{(f_x, f_y, 0) \mid 0 \leq f_x \leq 6 \wedge 0 \leq f_y \leq 6\} - R_{DC} - R_{L0}$, $R_{M1} = \{(f_x, f_y, 1) \mid 0 \leq f_x < 128 \wedge 0 \leq f_y < 64\} - R_{L1}$, and $R_{MM} = \{(f_x, f_y, f_t) \mid 0 \leq f_x < 128 \wedge 0 \leq f_y < 64 \wedge 2 \leq f_t \leq 4\}$. The high one is represented with $R_{H0} \cup R_{HH}$, where $R_{H0} = \{(f_x, f_y, 0) \mid 0 \leq f_x \leq 128 \wedge 0 \leq f_y \leq 64\} - R_{DC} - R_{L0} - R_{M0}$, and $R_{HH} = \{(f_x, f_y, f_t) \mid 0 \leq f_x < 128 \wedge 0 \leq f_y < 64 \wedge 5 \leq f_t \leq 8\}$.

The dissimilarity proposed is experimentally evaluated. We used two settings of dissimilarity: the normal and the customized settings. In the normal setting, all of the coefficients are set to one. In the customized setting, the coefficients are set as follows: $C_{DC} = 0.5, C_{l0} = 0.8$, $C_{lN} = 0.9, C_{m0} = 0.8, C_{mN} = 0.9, C_{h0} = 0, C_{hN} = 0.9$. The precision–recall curves for the key movie whose screen shot is shown in Fig. 1(a) are shown in Fig. 2. It is shown that the accuracy of the retrieval of the customized setting is good. The screen shot of the first candidate of the retrieval at the customized setting is shown in Fig. 1(b). This shot is very similar to that of the key movie shown in Fig. 1(a).

1) Hochin, T. *et al.* : Study on Quick Similarity Search in Massive-size Waveform Databases, Annual Report of National Institute for Fusion Science, April 2009 – March 2010, pg. 184 (2010).



(a) Key movie

(b) Retrieval result (1st candidate in customized setting)

Fig. 1: Key and result plasma movies.

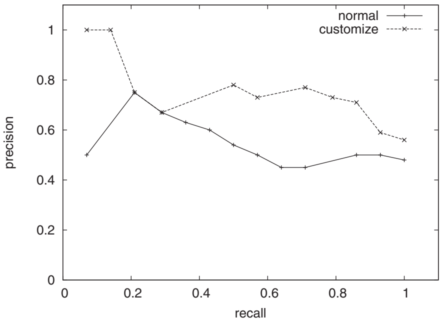


Fig. 2: Precision – recall curves.