Search and retrieval method of similar plasma waveforms

journal or publication title	Fusion Engineering and Design
volume	Vol.71
number	Issues 1-4
page range	pp.189-193
year	2004-06-01
URL	http://hdl.handle.net/10655/338

doi: 10.1016/j.fusengdes.2004.04.032

Search and Retrieval Method of Similar Plasma Waveforms

Nakanishi H.^{a,*}, Hochin T.^b, Kojima M.^a,
and LABCOM group^a

^a*National Institute for Fusion Science, 322-6 Oroshi-cho, Toki 509-5292, Japan*

^b*Osaka Prefecture University, 1-1 Gakuen-cho, Sakai 599-8531, Japan*

Abstract

Fusion plasma experiments often produce many channels of time-varying waveforms, and the assistance of a computational method to find similar waveforms could accelerate data analysis. To accelerate computer aided search, a so-called "thumb-indexed" reference will be required in the waveform database. Fourier series coefficients have been chosen for "R-tree" multidimensional indexing. As one waveform will be translated into one point in the multidimensional space of the Fourier series, the similarity between two points can be calculated as their Euclid distance in this space. For slowly varying waveforms like Large Helical Device (LHD) bolometer signals, a two-step search turned out to provide the best calculation speed of the relative distances, the first step uses the primary $(0, \pm 1, \pm 2, \dots, \pm k) = 2k + 1$ Fourier complex coefficients, and the second uses the other $2m$ ones whose amplitudes are the largest. Evaluation tests varying k and m have given the following best pair of values $2k + 1 = 5$ and $2m = 4$ for retrieving waveforms similar to those of the bolometer. The speed and precision of the method were also confirmed. This method makes the possibility to directly search similar phenomena in huge waveform databases realistic.

Key words:

LHD, bolometer, pattern matching, R-tree, Fourier coefficient, Euclidean distance

1 Introduction

Fusion plasma experiments often produce large quantities of time-dependent data. In particular, with steady-state experimental devices using superconducting mag-

* Corresponding author. phone: +81 572 58 2232 FAX: +81 572 58 2771
Email address: nakanishi@nifs.ac.jp (Nakanishi H.).

nets, the duration of the plasma becomes much longer, and therefore the size of the database increases drastically. In those cases, the assistance of a computational method will be useful to rapidly search and retrieve some specific data. To be able to find by numerical computation any waveforms similar to another one would accelerate various kinds of data analysis. In other words, a numerical method to quickly find out similar waveforms would be a new breakthrough [1].

Quick retrieval of similar waveforms among millions of data would be, however, very difficult if executed by one-by-one comparisons. To accelerate the computer-aided search, a so-called "thumb-indexed" reference will be also required in the waveform database.

So the objectives of this study are as follows: First, the method to find out similar waveforms by specifying a reference wave pattern must be defined. Second, some database analysis engine to accelerate the query search will become indispensable especially in the presence of huge amount of stored data. To reduce the mis-hitting ratio and obtain a better and faster search mechanism, some optimization of both the computational algorithm and the code implementation might be necessary for practical use.

2 Computation Method

In order to decide on the waveform searching method, the notion of waveform similarity must be defined first. A retrieval method to find similar time sequences has been proposed which uses Discrete Fourier Transformation (DFT) [2]. In this method, DFT is applied to extract the outward form characteristics of wave data and the $(0, \pm 1, \pm 2, \dots, \pm k)$ -th coefficients thus obtained are regarded as its characteristic values. In particular for slowly varying signals like plasma bolometer measurements which usually give trapezoidal waveforms, the lower DFT coefficients describe the outward form well.

As the waveform similarity can be understood as the similarity of their characteristic values, the next problem is to find how the amount of similarity between waveforms can be rewritten as a similarity between their characteristic parameters. The multidimensional Euclidean spatial length L between the two points $\vec{X}, \vec{Y}$ is used to determine their similarity.

$$L^2(\vec{X} - \vec{Y}) = |\vec{X}(0, \dots, k) - \vec{Y}(0, \dots, k)|^2 \quad (1)$$

With the DFT, a time series data $\vec{x} = [x_t] \ (t = 0, 1, \dots, n - 1)$ can be transformed

into $\vec{X} = [X_\omega]$ as follows:

$$X_\omega = \frac{1}{\sqrt{n}} \sum_{t=0}^{n-1} x_t e^{-i\omega t} \quad (2)$$

and its inverse transformation is also possible. In this way, the full series of DFT coefficients X_ω describes the original data completely. However, if the multi-D Euclidean distances are calculated in a space of large dimension k ,

$$|\vec{X} - \vec{Y}|^2 = \sum_{i=0}^k |X_i - Y_i|^2 \quad (3)$$

they will be very different from one another. This is a well-known fact called the curse of dimensionality [3]. Fig. 1 shows the corresponding conceptual mechanism. Therefore, a rough estimation is that using a number of indexes $k < 10$ will be good for most cases.

In order to construct the thumb-indexed reference for accelerating the query search among many database records, modern database management systems (DBMS) are usually equipped with a so-called "B-tree" (binary tree) indexing mechanism. It provides a 1D-index key only, and its multi-D extension is known as an "R-tree" (rectangle tree). Fig. 2 shows schematic views of B-tree and R-tree structures. Of course, R-trees can be extended not only to 2D but also to higher dimensions. The parent rectangular regions of every key points have cascading relations with outer container ones; for instance, the parent R1 contains a, b, and e at the bottom drawing of Fig. 2. Basically, the nearest neighbor points can be found by following down the cascade of the reference key point. However, in case of having an overlap in regions like R1 and R3 in Fig. 2, the subsequent tree search would diverged and the computing efficiency become worse.

As the Fourier coefficients are complex numbers except for the first order, storing the first $0 \sim k$ -th order coefficients results in storing a $(2k + 1)$ -dimensional point. So the Euclidean spatial distance L between every two points will be calculated in a $(2k + 1)$ -dimensional space.

3 Application for Bolometer Waveforms

This search and retrieval method to find similar waveforms has been applied to LHD bolometer measurements. As the bolometer gives slowly-varying trapezoidal waveforms, the lower frequency components of the $0 \sim k$ -th Fourier coefficients describes their outward form well. Table 1 shows a typical example of the real and

imaginary parts of the $0 \sim k$ -th DTF coefficients. The amplitudes of higher order ($k > 3$) DFT coefficients are much smaller than those of the lower ones.

As transient phases or rapid changes in plasma signals often characterize the whole waveform, $0 \sim k$ -th DFT values only could not distinguish the difference between plasma behaviors. In this study, we have used a two-step selection to obtain a better similarity with the bolometer waveforms. The first step selects the candidates by using an R-tree of $(2k + 1)$ -dimension as described above. In the second step, the candidates obtained from the first step will be examined again for similarity by using the other m DFT coefficients whose amplitudes as complex numbers are the largest.

To find the best values for k and m , evaluation tests have been done on 1000 LHD bolometer data by varying m from 1 to 4 for k from 1 to 3. Table 2 shows a typical result of this 2-step selection for the query waveform of LHD bolometer #28910. In this test, the waveforms displaying the best similarity to the target waveform were already known, and the order in which they were ranked in the answer list has been evaluated from their average ranking for every k and m pair. The pair $k = 2$ ($2k + 1 = 5$) and $m = 2$ ($2m = 4$) is the best one for retrieving waveforms similar to those of the bolometer, as shown in Fig. 3.

Fig. 4 shows an example of the retrieval of waveforms similar to those of the bolometer. The top one is the query waveform #28910, and the middle and bottom ones are the first and second candidates which were obtained by this retrieval method. Both are quite similar to the query one, however, the 2nd candidate #28911 seems to be more different for the start-up shoulder form and the peak height comparing to the 1st one. Their difference can be numerically evaluated as their distances of 8.60 and 9.05 from the query point, as shown in Table 2(b). These results prove the efficiency of the method very well.

4 Discussion

Computational speed will be also very important for any practical use of this waveform search and retrieval method. As for the R-tree index generation by using the 1000 bolometer data, it takes

3.660u 5.150s 0:14.19 62.0%

as user CPU seconds, system CPU seconds, real elapsed time, and the percentage of CPU occupation, respectively, on a Dual Pentium III 1.0GHz, 256 MB machine running on RedHat Linux 7.2 kernel 2.4.9. The generated index size is 248996 bytes.

To search the nearest neighbor 10 points in R-tree, it takes

177.910u 173.000s 5:56.59 98.4%

for all of the 1000 waveforms' queries. This means that one single search query takes about 0.36 second in average. All the implementation codes are written in Java version 1.3.1.

As the search is sufficiently rapid for 1000 waveforms, we can conclude that this method can be used in practical cases. Thus, the possibility to carry out a direct search for similar patterns in a huge waveform database becomes quite realistic.

DFT-based methods also have some open disadvantages. Their capability for slowly-varying plasma waveforms has been demonstrated here, however, other fluctuating signals having frequent zero-crossings present some difficulties to find a waveform similarity. In addition, this method has also a tendency to distinguish between waveforms whose lengths are slightly different and the key one excessively against human recognition. So, we have to note that this selection algorithm needs to be tuned for each diagnostic to prevent missing relevant matches.

Further research should be continued to improve the retrieval precision by applying other index trees like the SR-tree [4]. There is still room to improve the calculation speed by using other programming languages or applying faster DFT calculation routines.

References

- [1] T. Hochin, H. Nakanishi, M. Kojima, On the Construction of Databases of Experiment Data, NIFS Annual Report April 2001-March 2002, NIFS (2002).
- [2] D. Radiei, A. Mendelzon, Efficient Retrieval of Similar Time Sequences Using DFT, in: Proc. 5th Int'l Conf. on Foundations of Data Organization (FODO'98), 1998, pp. 249–257.
- [3] N. Katayama, S. Satoh, Indexing Methods for Similarity Retrieval, IPSJ Magazine 42 (10) (2001) 958–964, (in Japanese).
- [4] N. Katayama, S. Satoh, The SR-tree: An Index Structure for High-Dimensional Nearest Neighbor Queries, in: Proc. 1997 ACM SIGMOD Int'l Conf. on Management of Data, 13-15 May 1997, Tucson, Arizona, 1997.

List of Tables

1	Typical DFT coefficients of the LHD bolometer waveform: Re . and Im . correspond to the real and imaginary parts of the Fourier coefficients, respectively.	7
2	2-step nearest neighbor search: (a) First in terms of primary $2k + 1$ Fourier coefficients, and (b) second in terms of the largest $2m$ ones.	7

List of Figures

1	Conceptual view of the curse of dimensionality.	8
2	Schematic explanation of an R-tree: Typical result of the nearest neighbor query: (top) 1-D case, same as B-tree. (middle) 2-D case, and (bottom) 2-D case where some outer-container rectangles are overlapping.	9
3	Examination result to obtain the best k and m pair.	10
4	Typical result of the nearest neighbor query: (top) Key waveform. (middle) 1st, and (bottom) 2nd waveforms obtained in the ranking of Table 2(b). Units of axes are the number of samples in X, and the raw 12-bit ADC outputs of range $0 \sim 4095$ in Y, which is proportional to the signal voltages with some offsets.	11

Table 1

Typical DFT coefficients of the LHD bolometer waveform: *Re.* and *Im.* correspond to the real and imaginary parts of the Fourier coefficients, respectively.

k	0	1	2	3	4	5	6
<i>Re.</i>	2059	8	-7	-12	-9	-6	-4
<i>Im.</i>	NA	-14	-14	-4	0	1	1

Table 2

2-step nearest neighbor search: (a) First in terms of primary $2k + 1$ Fourier coefficients, and (b) second in terms of the largest $2m$ ones.

(a)						(b)		
data	DFT coefficients ($2k + 1$)					data	distance	rank
Bolo28910	2031	18	13	27	-19	Bolo28910	0.00	0
Bolo28948	2058	14	-7	10	-15	Bolo28917	8.60	1
Bolo28968	2064	17	-12	8	-20	Bolo28911	9.05	2
Bolo28949	2069	21	-12	12	-21	Bolo28916	40.53	3
Bolo28950	2071	24	-12	10	-27	Bolo28960	45.24	4
Bolo28967	2070	23	-13	11	-27	Bolo28943	61.77	5
Bolo28916	2021	10	23	28	-7	Bolo28968	62.41	6
Bolo28943	2027	30	17	29	-36	Bolo28967	68.11	7
Bolo28960	2014	20	30	29	-25	Bolo28948	68.42	8
Bolo28911	2035	18	12	23	-18	Bolo28950	68.69	9
Bolo28917	2029	14	11	22	-19	Bolo28949	75.93	10
Bolo28935	2042	30	2	7	-40	Bolo28935	82.70	11

Fig. 1. Conceptual view of the curse of dimensionality.

Fig. 2. Schematic explanation of an R-tree: Typical result of the nearest neighbor query: (top) 1-D case, same as B-tree. (middle) 2-D case, and (bottom) 2-D case where some outer-container rectangles are overlapping.

Fig. 3. Examination result to obtain the best k and m pair.

Fig. 4. Typical result of the nearest neighbor query: (top) Key waveform. (middle) 1st, and (bottom) 2nd waveforms obtained in the ranking of Table 2(b). Units of axes are the number of samples in X, and the raw 12-bit ADC outputs of range $0 \sim 4095$ in Y, which is proportional to the signal voltages with some offsets.