社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2006－DBS－139 (1)
2006－ＦＩ－ 83 (1)
2006／5／30

周波数に基づく波形の非類似度と類似検索への応用

宝珍 輝尚＊　小山 克正†　中西 秀哉‡　小嶋 護‡

＊京都工芸繊維大学 大学院工芸科学研究科 情報工学部門
〒606-8585 京都市左京区松ヶ崎御所海道町
hochin@kit.ac.jp

†大阪府立大学 大学院理学系研究科 情報数理科学専攻
〒599-8531 堺市学園町１－１

‡核融合科学研究所
〒509-5292 岐阜県土岐市下石町３２２－６
{nakanisi, kojima}@nifs.ac.jp

本論文では高周波成分の大きい波形の非類似度について検討する．高周波成分の大きい波形の比較においては，周波数領域の距離に基づく非類似度の方がうまく動作することを示す．この非類似度の良好性を，情報検索の評価に使用される適合率と再現率を用いて評価する．周波数領域の非類似度を時間領域の非類似度と比較した結果，周波数領域の非類似度の方が波形の検索において良好であることを明らかにする．また，本論文では，類似波形を検索する方法も提案する．提案する方法は周波数領域の非類似度を用いた方法である．この方法は，波形の部分検索に対しても利用可能である．

Frequency-Based Dissimilarity of Waveforms and Its Application to the Similarity Rerieval

Teruhisa HOCHIN＊　Katsumasa KOYAMA†　Hideya NAKANISHI‡　Mamoru KOJIMA‡

＊Kyoto Institute of Technology
Goshokaido-cho, Matsugasaki, Sakyo-ku, Kyoto-shi, Kyoto 606-8585 Japan
hochin@kit.ac.jp

†Osaka Prefecture University
1-1, Gakuen-cho, Sakai, Osaka, 599-8531 Japan

‡National Institute for Fusion Science
322-6, Oroshi-cho, Toki, Gifu, 509-5292 Japan
{nakanisi, kojima}@nifs.ac.jp

This paper studies on the dissimilarity of waveforms. It is shown that the dissimilarity based on the distance in the frequency domain works well in comparing the waveforms having major power at high frequency. The correctness of this dissimilarity is evaluated through the metrics used in evaluating that of the information retrieval, i.e. precision and recall. The frequency domain dissimilarity is compared with the time domain dissimilarity, which is the dissimilarity based on the distance in the time domain. The experimental result shows that the frequency domain dissimilarity works better than the time domain one in retrieving waveforms. This paper also proposes a new method of retrieving similar waveforms. The proposed method is based on the frequency domain dissimilarity. The method can be used for the subsequence matching of waveforms.

1 Introduction

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

Similarity retrieval of time series data (waveforms) have extensively been investigated [1-8, 10-22]. A waveform is represented with a point in a high dimensional space. The point is managed by using a multi-dimensional index structure. Because of the curse of dimensionality, reducing the dimension of waveforms is inevitable. The methods using the time-domain feature values have been proposed for the dimensionality reduction. These methods use Piecewise Aggregate Approximation (PAA)[12], Adaptive Piecewise Constant Approximation (APCA)[11], etc. The methods using the frequency-domain feature values have also been proposed. These methods use Discrete Fourier Transformation (DFT)[1, 19, 12, 17], Discrete Wavelet Transformation (DWT)[3, 10, 12], etc. These time-domain and frequency-domain feature values may be adopted for the purpose of the quick and correct retrieval of waveforms. It is considered that these feature values must reflect the characteristics of the dissimilarity of waveforms. If not, the retrieval result will contain many false drops. These are not appropriate for the feature value. It is, therefore, important to clarify the dissimilarity assumed.

Major dissimilarity of waveforms is the Euclid distance of time series composing the waveforms. The dissimilarity is the fundamental and important measure in retrieving waveforms because it is used to decide whether a waveform is similar to a key waveform. The Euclid distance of time series works well for the waveforms, of which changes are gradual and/or slow. It is, however, doubtful that it works well for the waveforms, of which changes are severe and/or quick. In this case, even if a waveform is a slightly shifted one to a key waveform, the Euclid distance may become large. This waveform could not remain in the query result. As this waveform should be in the query result, this means false dismissals.

Agrawal *et al* have proposed the similarity model allowing the translation and the amplitude scaling[2]. Rafiei *et al* have proposed the method supporting the wide variety of comparisons[20]. Kalpakis *et al* have defined the similarity based on the underlying physical models[8]. Although the distance may be domain-specific as described in [14], it is considered to be important to find the general distance used in a variety of areas in th point of view of the database research, where general solutions are required.

Investigating the dissimilarity of waveforms is important by the reasons described above. The dissimilarity is the fundamental and important measure in the similarity retrieval of waveforms. When the dissimilarity changes, the method retrieving waveforms efficiently may change.

This paper studies on the dissimilarity of a class of waveform. The major characteristics of the waveforms in the class is that the change is severe and/or quick. In other words, the waveforms have the dominant power at high frequency. This paper proposes the dissimilarity of this kind of waveform. The proposed distance is of the frequency domain. The correctness of the dissimilarity is evaluated through precision and recall, which are the metrics used in the evaluation of the information retrieval. It is shown that the frequency domain dissimilarity works better in comparing waveforms than the time domain one. Moreover, this paper proposes a method of retrieving similar waveforms based on the frequency domain dissimilarity.

Remaining of this paper is as follows. Section 2 proposes a frequency domain dissimilarity, which is called *spectrum distance*, after the definitions of the terms describing the dissimilarity. The frequency domain dissimilarity and the time domain one are evaluated in Section 3. In Section 4, a method of retrieving similar waveforms based on the frequency domain dissimilarity is proposed, and is evaluated. Section 5 discusses the dissimilarity and the proposed retrieval method. Finally, Section 6 concludes the paper.

2 Spectrum distance

The terms used in this paper are described before proposing a frequency domain dissimilarity.

The first is the Euclid distance. Given two sequences $x = x_1, \cdots, x_n$ and $y = y_1, \cdots, y_n$, their Euclidian distance is usually defined by Eq. (1).

$$D(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \qquad (1)$$

We use a variation of the Euclidian distance cal-

culated by Eq. (2).

$$D_t(x,y) = \sqrt{\sum_{i=1}^{n}((x_i - x_{ave}) - (y_i - y_{ave}))^2} \quad (2)$$

where x_{ave} and y_{ave} are the average values of the time series x and y, respectively. The major reason of subtracting the average value from the original values is that the original values may include an offset value because of the calibration adjustment.

Next is the Discrete Fourier Transformation (DFT). The n-point DFT of a signal $x = x_1, \cdots, x_n$ is defined as a sequence X of n complex numbers X_f defined by Eq. (3).

$$X_f = \frac{1}{\sqrt{n}} \sum_{i=0}^{n-1} x_i \exp(2\pi i f t/n), f = 0, 1, \cdots, n-1 \quad (3)$$

The value X_f is a complex value. The real part of X_f brings the amplitude information, and the imaginary part brings the information on the phase.

Finally, we propose the *spectrum distance* as the dissimilarity of waveforms. This distance uses coefficients of the Discrete Fourier Transformation. Given two Discrete Fourier Transforms X_f and Y_f, their spectrum distance is defined by Eq. (4).

$$D_s(x,y) = \sqrt{\sum_{i=1}^{n-1}(|X_f| - |Y_f|)^2} \quad (4)$$

The absolute value $|X_f|$ is called the *FFT component*. This corresponds to the power information. Please note that the summation begins at one rather than zero. As the 0th DFT coefficient corresponds to the average value of a time series, we omit the 0th DFT coefficient by the reason described at the explanation of Eq. (2).

From here on, the Euclid distance defined by Eq. (2) is called the *time-ordering distance* for distinguishing it from the spectrum distance.

3 Evaluation of the distance

The spectrum distance is evaluated from the point of view of the accuracy. The evaluation is based on the measures of evaluating the accuracy of information retrieval: *precision* and *recall*.

The waveforms obtained through the fusion plasma experiments are used in the evaluation. Examples of the waveforms are shown in Fig. 1. The waveform is originally consisted of about 130,000 points of data. It is too many to be used in these

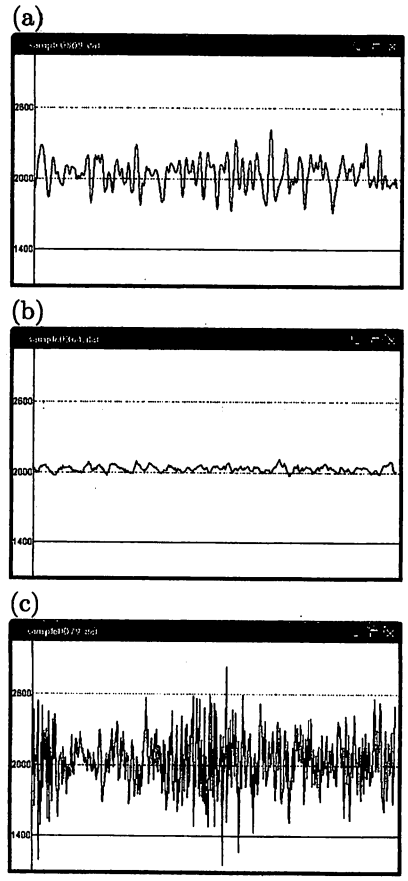

Figure 1: Key waveforms

experiments. A series consisted of 512 points is derived from an original waveform. The number of the waveforms used is 1000.

One of the authors decided which waveforms are similar to the key waveforms. The waveforms obtained are called the correct waveforms to a key waveform.

The spectrum and the time-ordering distances between a key waveform and the 1000 waveforms are calculated. The waveforms are ordered according to the distances obtained. Precision and recall are calculated by using the order of the waveforms. Precision is the ratio of the number of the correct answers within the result to the number of the result. Recall is the ratio of the number of the correct answers within the result to the number of all correct answers.

The three waveforms shown in Fig. 1 are used as the key ones in the evaluation. These are different from one another. The amplitude of the

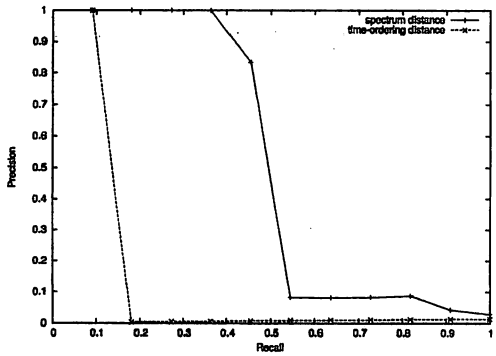

Figure 2: Evaluation result for Key (a)

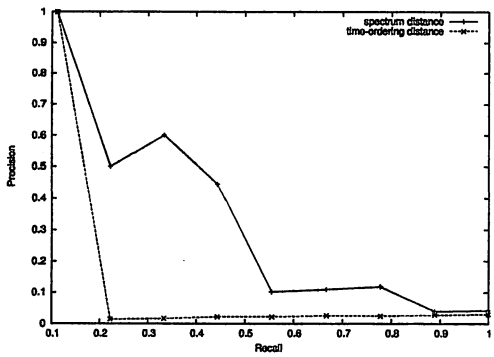

Figure 3: Evaluation result for Key (b)

key waveform (b) is less than the waveform (a). On the other hand, that of Key (c) is larger than Key (a).

The results of the evaluations are shown in Fig. 2 to Fig. 4. The key waveform of the evaluation result shown in Fig. 2 (Fig. 3 and Fig. 4) is Fig. 1(a) ((b) and (c), respectively). The key waveform is included in the 1000 waveforms. The first candidates of all results were the key waveform themselves. Therefore, the curves begin at 1.0 on the precision axis. In all of the evaluation results, the spectrum distance has better characteristics than the time-ordering one. The second candidates obtained according to the time-ordering distance were not in the correct waveforms in all of the evaluation results. This fact degrades the accuracy of the time-ordering distance. The first several candidates obtained by using the spectrum distance belong to the correct waveforms. This is the reason that the spectrum distance has

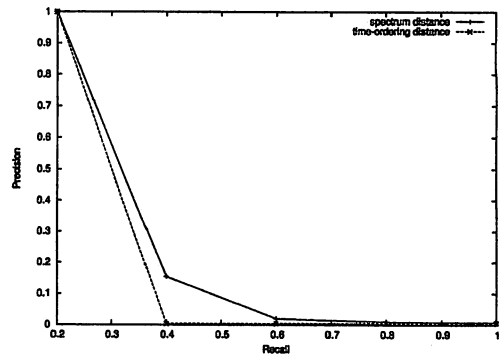

Figure 4: Evaluation result for Key (c)

the good characteristics as the dissimilarity measure. It is considered that the spectrum distance is better than the time-ordering one as the dissimilarity of this type of waveform.

4 Retrieval method based on the spectrum distance

Here, we propose a simple method retrieving similar waveforms. The proposed method is based on the spectrum distance.

4.1 Index construction

A waveform is divided into m segments. Each segment has n points, where n is the power of two because the Fast Fourier Transformation (FFT) is applied to a segment for obtaining the frequency components. For every segment, the following procedure is applied to. FFT components are obtained for a segment. FFT components are divided into k segments, which are called *frequency segments*. An average value is calculated for each frequency segment. As FFT components are divided into k frequency segments, k average values are obtained. These average values are inserted into a multi-dimensional index as a point in the k dimensional space. A point has the shot number and the segment number as its attribute values. The shot number is the number distinguishing a waveform from the other ones. The segment number is the serial number put to a segment in a waveform. Moreover, an additional file, which is called the *seek file*, is used for the quick retrieval. This file contains the segment information. An entry of the file contains the k average values of a

```
(1) Divide a waveform into m segments,
    each of which includes n points
(2) Repeat the followings for each of
    m segments
  (2-1) Obtain FFT components from
        a segment
  (2-2) Divide FFT components into
        k frequency segments
  (2-3) Calculate k average values
        for the frequency segments
  (2-4) Insert the k-dimensional point
        having k average values into
        a multi-dimensional index
  (2-5) Insert k average values and
        a segment information into
        a seek file
```

Figure 5: Procedure of index construction

segment. The entries are placed according to the order of the shot number and the segment number. If a shot is missing, the entries are filled with values of zero. Please note that the next segment of a segment is placed next to it. As a key waveform usually contains several segments, we have to evaluate a series of segments. When a segment may be the one in a candidate waveform, the other segments that should be evaluated can easily be obtained by using the position because the segments lie in order. The seek file is used for this purpose. As it decreases the time of looking the index up, retrieval may have good performance. The outline of registration of a waveform is shown in Fig 5.

4.2 Retrieving waveforms

When a key waveform is given, it is divided into m segments. FFT is applied to each segment to obtain FFT components, and k average values of its frequency segments. By using k average values of the ith segment of a waveform, where $i = 1, \cdots, m$, the multi-dimensional index is looked up to obtain the candidate segments in the specified region of multi-dimentional space. Next, the seek file is consulted. The remaining $m - 1$ segments of a waveform are obtained by using the segment found in the above step as the ith segment. The k average values are obtained for each segment. The Euclid distance of m segments is calculated. This Euclid distance is called the *index distance.* After the index distances are calculated for all of the waveforms obtained, those having large index distance are discarded. FFT is applied to the segments of the remaining waveforms, and the spectrum distances are calculated. Finally, the waveforms are sorted according to the order of the spectrum distance. The outline of retrieving waveforms is shown in Fig 6.

```
(1) Divide a key waveform into m seg-
    ments, each of which includes n
    points
(2) For each of m segments, look up
    the multi-dimensional index to
    obtain candidate segments
(3) Obtain waveforms, each of which
    includes a candidate segment
(4) Calculate the index distance
(5) Discard the waveforms having large
    index distance
(6) Calculate the spectrum distance
(7) Sort the series of waveforms
```

Figure 6: Procedure of retrieval

4.3 Evaluation

The proposed retrieval method is evaluated. The 10,000 waveforms obtained through the fusion plasma experiments are used. The waveform is originally consisted of about 130,000 points of data. A waveform is divided into 256 segments. Each segment has 512 points. There are 256 * 10,000 segments in a database. FFT components are divided into 4 frequency segments. The SR-tree[9] is used as the multi-dimensional index structure. As the number of the frequency segments is four, a segment is stored as a point in a four-dimensional space. The size of the SR-tree is about 500 MB. That of the seek file is 39MB.

The key waveform and the waveforms obtained are shown in Fig. 7. This key waveform has 2048 points. It is divided into four segments ($k = 4$). The segment of the query range of the multi-dimensional index is [-0.12, 0.12] for every dimension. The value 0.12 is decided experimentally according to the pre-experiment of the retrieval. The threshold value of discarding the waveforms having the large index distance is $0.24 * k$. The value 0.24 is also experimentally decided accord-

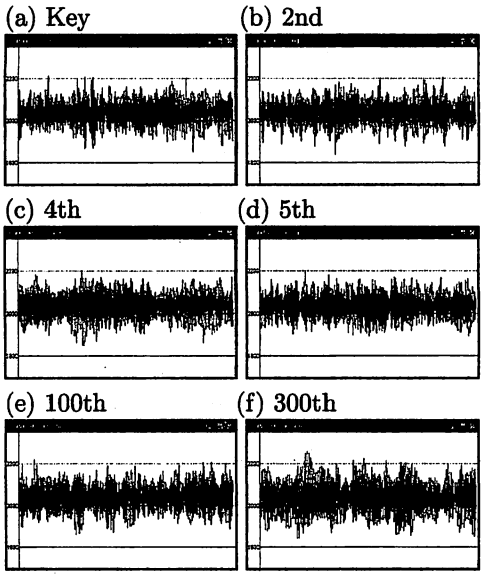

Figure 7: The key and the several waveforms obtained (1).

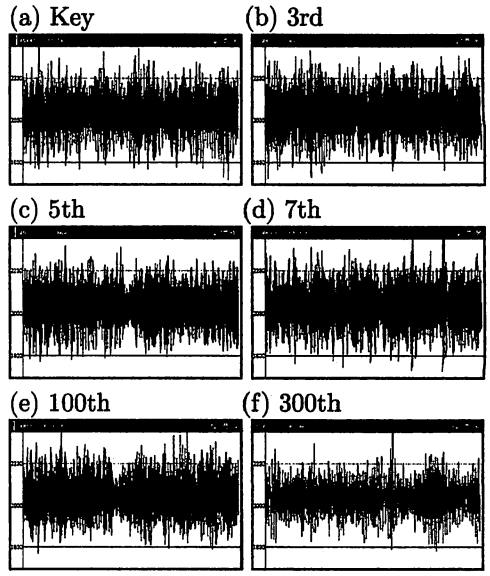

Figure 8: The key and the several waveforms obtained (2).

ing to the pre-experiment of the retrieval. The number of the final candidates of the retrieval is 307. In Fig. 7, the second, the fourth, the fifth, the hundredth, and the three hundredth candidates are shown. The third one is not shown because it is a slightly different series of the key waveform, and it is very similar to the key one. It seems that the waveforms obtained are similar to the key waveform. Another result is shown in Fig. 8. The key waveform has 8192 points, and is divided into 16 segments. The second, the fourth, and the sixth, and so on are not shown because of the similar reason in the previous result. It also seems that the waveforms obtained are similar to the key one.

The retrieval time is measured by varying the number of the candidates. The number of the candidates are changed by using different key waveforms. Every waveform has 16 segments. In this case, it takes 16 hours to create the index, and creating the seek file takes 20 minutes. The experimental programs run under Windows XP operating system on a custom-made personal computer composed of Intel Celeron 2.5GHz processor, 256MB memory, 60GB HDD, of which rotation speed is 5400 rpm.

The result is shown in Fig. 9. The index look-up time, the seek file access time, and the calculating spectrum time as well as the total time are shown. When the number of candidates is large, the time of calculating spectrum distance is dominant. On the other hand, the index look-up time and the seek file access time are even. Decreasing the number of candidates seems to be effective for the performance of retrieval.

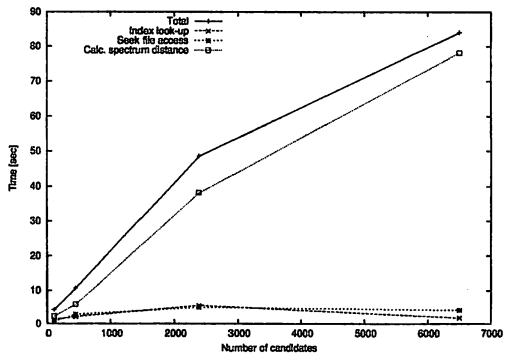

Figure 9: Retrieval time and the number of candidates

5 Discussion

It has been shown that the first several FFT coefficients are the good feature values for the waveforms, of which changes are slow[1, 19]. The absolute values of the first several FFT coefficients of these waveforms are dominantly large. In this case, these coefficients are good enough to represent the feature of the waveforms. These are considered to be a kind of frequency domain dissimilarity. The severely-changing waveforms are mainly treated in this paper. The low frequency FFT coefficients do not work well as the feature values for this type of waveform because the peak values of the FFT coefficients appear at the high frequency. The spectrum distance considers this kind of waveform. The high frequency coefficients are taken into account to the dissimilarity.

The time domain feature values, e.g. PAA[12] and APCA[11], approximate waveforms by representing them with the lower frequency step functions. It is considered that these capture the low frequency tendency of waveforms. This means that the high frequency components of a waveform are discarded.

The proposed retrieval method is developed for the subsequence matching to waveforms. We have demonstrated that a series of several segments is searched in the waveforms, each of which has 256 segments, in the performance evaluation of the proposed retrieval method.

The retrieval procedure is consisted of two major phases. The first one is the pruning phase. A little bit large number of data are selected from a database. The second phase is the checking one. The selected data is examined whether they can remain in the query result. In this phase, the distance is calculated by using all of the points in a series of segments. This is a time-consuming task as we have seen in the performance evaluation. If all of the FFT coefficients of all of the segments are stored in a database or file, this could accelerate the speed in calculating the distance. However, intolerably huge spatial cost will be required. Balancing the trade-off between the performance cost and the spatial one is in the future work.

The proposed method uses a file separated from a multi-dimensional index. This file is called the seek file. Introducing the seek file could cause the decrease of the time of looking up the index as mentioned before. Another reason of adopting the seek file is that we could use the program for the multi-dimensional index structure developed by the other researchers and/or developers. The methods without using the seek file include the method obtained by modifying the program for the multi-dimensional index structure in order to make the sequential access of segments possible. This may be another solution.

6 Conclusion

This paper experimentally clarifies that the spectrum distance works well as the dissimilarity of waveforms of which changes are severe and/or quick. The spectrum distance is the distance in the frequency domain. This distance is evaluated by using precision and recall. A method of retrieving similar waveforms is also proposed. A waveform is divided into several segments. FFT is applied to a segment. FFT components are divided into several segments (frequency segments). A feature value is calculated for each frequency segment. We adopt an average value of the FFT components in a frequency segment as the feature value. A segment is managed as a point in a high dimensional space by using a multi-dimensional index. The feature values of all of the segments are also stored in a file (seek file) according to the order of the segments. Owing to this file, the time of accessing the index can be decreased. The retrieval experiments show that the similar waveforms can be obtained, and decreasing the number of the candidates retrieved may be effective for the retrieval performance.

The spectrum distance works well for the waveforms described above. It may not, however, work for the waveforms, of which changes are slow. For this type of waveform, the phase information as well as the amplitude one have to be considered. The application of the spectrum distance to the slowly changing waveforms is one of the future works. The proposed retrieval method divides a waveform into several segments. The position of dividing a waveform may cause another type of dismissals. Revealing the relationships between the position of the division and the accuracy of the retrieval is also in the future works.

References

[1] R. Agrawal, C. Faloutsos, and A. N. Swami. Efficient similarity search in sequence databases. In *FODO*, pages 69–84, 1993.

[2] R. Agrawal, K.-I. Lin, H. S. Sawhney, and K. Shim. Fast similarity search in the presence of noise, scaling, and translation in

time-series databases. In *VLDB'85*, pages 490–501, 1995.

[3] K.-P. Chan and A. W.-C. Fu. Efficient time series matching by wavelets. In *15th ICDE*, pages 126–133, 1999.

[4] C. Faloutsos, M. Ranganathan, and Y. Manolopoulos. Fast subsequence matching in time-series databases. In R. T. Snodgrass and M. Winslett, editors, *1994 ACM SIGMOD Conference*, pages 419–429, 1994.

[5] L. Harada. An efficient sliding window algorithm for detection of sequential pattern. In *DASFAA'03*, pages 73–80, 2003.

[6] S. Hirano and S. Tsumoto. Mining similar temporal patterns in long time-series data and its application to medicine. In *ICDM*, pages 219–226, 2002.

[7] T. Kahveci and A. K. Singh. Optimizing similarity search for arbitrary length time series queries. *IEEE Trans. Knowl. Data Eng.*, 16(4):418–433, 2004.

[8] K. Kalpakis, D. Gada, and V. Puttagunta. Distance measures for effective clustering of arima time-series. In *ICDM*, pages 273–280, 2001.

[9] N. Katayama and S. Satoh. The sr-tree: An index structure for high-dimensional nearest neighbor queries. In *1997 ACM SIGMOD Conference*, pages 369–380, 1997.

[10] K. Kawagoe and T. Ueda. A similarity search method of time series data with combination of fourier and wavelet transforms. In *TIME*, pages 86–92, 2002.

[11] E. J. Keogh, K. Chakrabarti, S. Mehrotra, and M. J. Pazzani. Locally adaptive dimensionality reduction for indexing large time series databases. In *2001 ACM SIGMOD Conference*, 2001.

[12] E. J. Keogh, K. Chakrabarti, M. J. Pazzani, and S. Mehrotra. Dimensionality reduction for fast similarity search in large time series databases. *Knowl. Inf. Syst.*, 3(3):263–286, 2001.

[13] Y. Kim, Y. Park, and J. Chun. A dynamic indexing structure for searching time-series patterns. In *COMPSAC*, pages 270–275, 1996.

[14] Q. Li, I. F. V. López, and B. Moon. Skyline index for time series data. *IEEE Trans. Knowl. Data Eng.*, 16(6):669–684, 2004.

[15] W.-K. Loh and S.-W. Kim. A subsequence matching algorithm supporting moving average transform of arbitrary order in time-series databases using index interpolation. In *ADC*, pages 37–44, 2001.

[16] W.-K. Loh, S.-W. Kim, and K.-Y. Whang. Index interpolation: An approach to subsequence matching supporting normalization transform in time-series databases. In *CIKM*, pages 480–487, 2000.

[17] H. Nakanishi, T. Hochin, M. Kojima, and L. group. Search and retrieval method of similar plasma waveforms. *Fusion Engineering and Design*, 71:189–193, 2004.

[18] S. Park, D. Lee, and W. W. Chu. Fast retrieval of similar subsequences in long sequence databases. In *1999 Workshop on Knowledge and Data Engineering Exchange*, pages 60–67, 1999.

[19] D. Rafiei and A. O. Mendelzon. Efficient retrieval of similar time sequences using dft. In *FODO'98*, pages 249–257, 1998.

[20] D. Rafiei and A. O. Mendelzon. Querying time series data based on similarity. *IEEE Trans. Knowl. Data Eng.*, 12(5):675–693, 2000.

[21] P. Seshadri, M. Livny, and R. Ramakrishnan. Sequence query processing. In *1994 ACM SIGMOD Conference*, pages 430–441, 1994.

[22] C. Shahabi, X. Tian, and W. Zhao. Tsa-tree: A wavelet-based approach to improve the efficiency of multi-level surprise and trend queries on time-series data. In *SSDBM2000*, pages 55–68, 2000.