ISSN (Online): 1694-0784
ISSN (Print): 1694-0814

Fast Clustering of Self-Similar Network Traffic Using Wavelet

Aqil Burney S.M.[1], Akhter Raza Syed[2], and Afzal Saleemi[3]

**[1] Chairman at Department of Computer Science, University of Karachi
Karachi, Pakistan**

**[2] Research fellow at Department of Computer Science, University of Karachi
Karachi, Pakistan**

**[3] Head Department of Computer Science, National University of Sciences and Technology
PMA Kakul, Pakistan**

Abstract

In this paper we have given our proposed approximation methods for similarity search in large temporal databases using wavelet transformation based featured signals and time warping distance algorithm. Our main goal is to propose efficient methods to speed up the mining of matched sequences especially when the sequences are of random lengths and traditional distance metrics like Euclidean distance fail to achieve the desire goals. We proposed two methods for truncation of databases for optimizing search procedures for similarities using the concept of wavelet based featured time warping. In our first model, we utilize the maxima, minima and average features of wavelet based compressed signals and in second model, features of wavelet transformation using average of approximation coefficients at the coarsest scale and maxima of maxima and minima of minima of detail coefficients at all scales. We show by carrying out extensive experiments that our proposed methods are very effective and ensure the nonoccurrence of false dismissals and minimal false alarms with least compromise over accuracy.

Keywords: *Wavelets, Multiresolution analysis, Dimensionality reduction, Network Traffic, Time warping, Data mining*

1. Introduction

During the last few years, explosion of information has created extremely large databases and is piling up large mountains of data on daily basis. At the same time highly sophisticated machines with immense computational facilities have made the things easier for data manipulators and decision makers to obtain optimum information out of this data flood. This also gives emergence to a new field of study called data mining. Data mining is defined to mine out information from huge databases for end users to make use of it for decision making process. In the field of data mining, the efficient retrieval of time based information hidden in mountains of data is of great importance.

In past few years due to high speed computers, time series data has once again gained importance in financial analysis, marketing, data mining, data warehousing, geology, computer and network engineering etc. With this growing demand of time series data, there is increasing demand to support fast retrieval of time series data based on similarity measurement. In order to find similarity between two time series, we have to define similarity measurements during the query process. There are number of distance metrics like Euclidean, Mahalanobis, cityblock, Minkowski, cosine, correlation and many others which have been used for similarity search in previous literature [1] [5]. In this chapter we use the dynamic time warping distance measure, which we discuss at later stages. The efficient and fast retrieval of similar time series in huge databases is only possible with feature extraction or dimension reduction of data [2] [3] [4]. There are number of transformations available like singular value decomposition (SVD) [6], piecewise aggregate approximation (PAA), fast Fourier transformation (FFT) for dimension reduction [7] [8]. We used discrete wavelet transformation (DWT) for compression and feature extraction due to its added advantages over contemporary transformations [9] [10].

The remainder of this paper is structured as follows in the next section, we describe how the data can be decomposed into multi-resolution levels using a robust smoother-cleaner DWT and reconstructed with outlier patches removed using dimension reduction technique. Section 3 describes the dynamic time warping technique. In section 4, we give our proposed technique and perform extensive experiments on simulated synthetic self similar network traffic signals using wavelets and time warping for arbitrary length signals supporting index based techniques. Section 5 contains our concluding remarks.

2. Dimensionality Reduction using Wavelets

There are various techniques, which are being used for dimension reduction. The dimension of a finite dimensional vector space 'V', denoted by $\dim(V)$ is defined to be the number of vectors in a basis for 'V' where 'V' is any vector space. The basic concept of basis function is that if 'V' is any vector space and $S = \{v_1, v_2 \cdots v_n\}$ is a set of vectors in 'V', then 'S' is called a basis for 'V' if the following conditions hold

(a) 'S' is linearly independent i.e. $k_1 v_1 + k_2 v_2 + \cdots k_r v_r = 0$ where $k_1, k_2 \cdots k_r$ are scalar quantities.
(b) 'S' spans 'V'

The basic idea behind the dimension reduction is that in most types of data the energy of the time series signal is concentrated in some part of it and rest of the signal does not significantly contribute towards energy and considered as noise. Transforming each sequence of data and keeping only subset of transformed coefficients performs the dimension reduction.

Briefly, using wavelet transforms, a network traffic signal can be decomposed into a cost-effective countable set of basis functions at different time locations and resolution levels. Unlike Fourier analysis, wavelet analysis captures the more localized behavior in a signal. Trigonometric functions serve as functions on which a Fourier decomposition of time series data is based in the frequency domain. In contrast, wavelet analysis is characterized by basis functions that are not trigonometric and that have their energy concentrated within a short interval of time. These 'small waves', or wavelets, are defined over the square integrable functional space, $L^2(R)$, and they have compact support. It is the property of compact support that enables wavelet analysis to capture the short-lived, often momentary components of data that occur in shorter time intervals [11] [12]. Wavelets belong to families which provide the building blocks for wavelet analysis. Just as sine and cosine functions are functional bases onto which we project data to extract information belonging to the frequency domain, wavelet functions are functional bases that allow for extraction of information available in both the time and frequency domains. A wavelet family comes in pairs, a father and mother wavelet. The father wavelet $\varphi(t)$ represents the smooth and low frequency part of the signal, while the mother wavelet $\psi(t)$ captures the detail or high-frequency component.

A continuous function $f(t)$ can be approximated by the orthogonal wavelet series given by

$$f(t) \cong \sum_k s_{J,k}\varphi_{J,k}(t) + \sum_k d_{J,k}\psi_{J,k}(t) + \sum_k d_{J-1,k}\psi_{J-1,k}(t) + \cdots \sum_k d_{1,k}\psi_{1,k}(t) \quad (1)$$

where J is the number of multi-resolution components or scales and k ranges from one to the number of coefficients in a multi-resolution component.

The coefficients $s_{J,k}, d_{J,k}, d_{J-1,k} \cdots d_{1,k}$ are the wavelet transform coefficients, while $\varphi_{J,k}(t)$ and $\psi_{J,k}(t)$ are the approximating father and mother wavelet functions respectively. The wavelet approximation to $f(t)$ given by Eq. (1) is orthogonal since the basis functions φ and ψ are orthogonal by construction [11] [12]. Wavelet functions usually do not have a closed functional form. After imposing desired mathematical properties and characteristics, they are generated through dilation and translation according to the following normalized functions.

$$\varphi_{j,k}(t) = 2^{-\frac{j}{2}}\varphi\left(\frac{t - 2^j k}{2^j}\right) \quad (2)$$

$$\psi_{j,k}(t) = 2^{-\frac{j}{2}}\psi\left(\frac{t - 2^j k}{2^j}\right) \quad (3)$$

The wavelet transform coefficients measure the contribution of the corresponding wavelet function to the approximating sum. Consider the set of father wavelet functions $\varphi(t)$, which span the sub-space V_J of $L^2(R)$,

$$V_J = \overline{Span\{\varphi_k(t)\}}$$

where

$$\varphi_k(t) = \varphi(t - k), \quad k \in Z \quad (4)$$

It follows that any function in the V_J space can be expressed as a linear combination of the father wavelets $\varphi_k(t)$, which span the space. That is

$$f(t) = \sum_k a_k \varphi_k(t), \quad \forall\ f(t)\ \varepsilon\ V_J \quad (5)$$

If a set of signals based on information set that represents the fundamentals can be expressed by the weighted sum given by (5), then a set of signals based on more detailed information set should be contained in a sub-space, V_j which contains V_J. The detail or higher frequency components of the signal are captured by the mother wavelets at higher levels of resolution. The subscript j that we incorporate into the mother and father wavelets represents the level of time resolution and is known as the dilation parameter [12]. Recall Eq. (2) for the father wavelets where j is the dilation parameter and k is the translation parameter that ensures the father wavelets span the V_j space. For the mother wavelets, Eq. (3) captures the extra detail over and above that accounted for by the father wavelets at a particular scale or dilation.

2.1 Multiresolution Analysis using Wavelets

The multiresolution condition requires that

$$V_j \supset V_{j+1} \quad \forall \; j \in Z$$
$$V_{-\infty} = \{0\}$$
$$V_{\infty} = L^2(R)$$

with the orthogonal complement of V_j in V_{j-1} being the subspace, W_j. W_j is spanned by orthogonal mother wavelet functions such that

$$V_{j-1} = V_j \oplus W_j \qquad \text{and}$$
$$L^2(R) = V_J \oplus W_J \oplus W_{J-1} \oplus W_{J-2} \oplus \cdots \oplus W_1$$

For a discrete signal $f = (f_1, f_2, \cdots f_n)'$ sampled from a continuous time signal $f(t)$, the discrete wavelet transform maps the vector f into a set of wavelet coefficients W , which contains the coefficients $s_{J,k}$ and $d_{j,k}$, $j = 1,2,\cdots J$. When the number of observations n is divisible by 2^J then the number of coefficients at any particular scale depends on the width of the wavelet function [11] [12] [13] [14]. At the finest scale 21, $\frac{n}{2^1}$ coefficients and for coarsest scale 2^J, $\frac{n}{2^J}$ coefficients are required. As the level of resolution descends to the smoothest level 2^J , the number of coefficients required decreases each time by a factor of 2. From the orthogonal property of wavelet transforms, it follows that

$$n = \frac{n}{2} + \frac{n}{2^2} + \cdots \frac{n}{2^{J-1}} + \frac{n}{2^J}$$

The detail coefficients $d_{J,k}$ give the coarse scale deviations from the smooth behavior at scale 2^J, which is represented by the smooth coefficients. The remaining detail coefficients $d_{J-1,k}, d_{J-2,k} \cdots d_{1,k}$ capture the progressively finer scale deviations from the smooth behavior. At a particular level of time resolution j the impact of the information subset on the signal is reflected in the number and magnitude of the wavelet coefficients and is roughly equal to the sampling interval at that resolution level. Information corresponding to finer detail in the signal than that at resolution level j can only be incorporated into the signal by considering shorter sampling intervals which are associated with higher levels of resolution than j . Such information will not contribute to approximating the signal at lower levels.

The terms of Eq. (1) are comprised of functions called the smooth signal, $S_J(t) = \sum_k s_{J,k}\varphi_{J,k}(t)$ and the detail signals, $D_j(t) = \sum_k d_{j,k}\psi_{j,k}(t)$ such that the orthogonal wavelet series approximation to $f(t)$ is

$$f(t) \cong S_J(t) + D_J(t) + D_{J-1}(t) + \cdots D_1(t) \qquad (6)$$

Eq. (6) is known as a multiresolution decomposition of $f(t)$ because the terms of different scales represent the components of the signal at different resolutions. Just as $V_J, W_J, W_{J-1}, \cdots\cdots\cdots W_1$ can be seen as a partition of the information set, information decomposition in Eq. (6) allows us to reconstruct the signal $f(t)$ based on a subset of relevant information at the j^{th} level of resolution, via the approximation,

$$S_{j-1}(t) = S_J(t) + D_J(t) + D_{J-1}(t) + \cdots D_j(t)$$

These approximations range from the smoothest scale or lowest level of resolution 2^J to finer scales $2^{J-1}, 2^{J-2}, \cdots 2$. Using the different multiresolution approximations $S_1(t), S_2(t), \cdots S_J(t)$, we focus on different features of the signal. The finer scale approximations reveal more details as a result of incorporating higher frequency observations and shorter time intervals between observations [12] [14]. In this experiment we use the wavelets for transformation of data using Eq. number (6) and global threshold methods for compression and significant features extraction. We utilize the features extracted from wavelet based synthesized compressed signal and the wavelet based transformed coefficients on the bases of their maxima, minima and average for our two proposed methods.

3. Dynamic Time Warping

There are numbers of distance metrics given in section 1, which are available for detecting similarities among series in database. In most of the cases, two network traffic signals have the same over all shape but do not align in X-axis. In such cases distance metrics given in section 1 do not meet the requirements. In order to find similarities between such self-similar traffic we have to align time axis of one series or both using the warping technique for obtaining better comparison [15] [16] [17]. Dynamic time warping (DTW) is the best possible technique to obtain such type of warping in two time series. DTW is extensively used in data mining, speech recognition, robotics, manufacturing, genetics, network traffic modeling and medicines [18] [19] [20] & [21].

3.1 Dynamic Time Warping (DTW) Algorithm

The classical DTW algorithm is given in figure1. Suppose we have two time series

$$\vec{x} = \{x_0, x_1.x_2, \cdots, x_{n-1}\}$$

and $$\vec{y} = \{y_0, y_1, y_2, \cdots, y_{m-1}\}$$

where $\|x\|$ and $\|y\|$ denote the length of vectors, not necessarily equal. $C_{matrix}[i][j]$ stores the shortest cumulative distance from $\{x_0, x_1, \cdots, x_i\}$ to $\{y_0, y_1, \cdots, y_j\}$ which is the Euclidean distance between pair (x_i, y_j) plus $C_{matrix}[i-1][j]$, $C_{matrix}[i][j-1]$ and $C_{matrix}[i-1][j-1]$. The time warping distance will be $D_{TW}(\vec{x}, \vec{y}) = [C_{matrix}[x_{n-1}, y_{m-1}]]^2$ and the shortest cumulative distance for each pair of two time series starts from pair (x_0, y_0) to (x_{n-1}, y_{m-1}) will be time warping path in $n \times m$ matrix.

Algorithm 1: Dynamic Time Warping Distance

Input: $\vec{x}, \vec{y}$ Output: C_{matrix}

1. $x_{len} = \|\vec{x}\|$; $y_{len} = \|\vec{y}\|$; $C_{matrix}[0][0] = 0.0;$
3. for $(0 \le i \le x_{len-1})$ $C_{matrix}[i][0] = \infty;$
4. for $(0 \le j \le y_{len-1})$ $C_{matrix}[0][j] = \infty;$
5. for $(0 \le i \le x_{len-1})$
6. for $(0 \le j \le y_{len-1})$;
7. $C_{matrix}[i][j] = \{(D(x_i, y_j))^2 + (\min(C_{matrix}[i-1][j], C_{matrix}[i][j-1], C_{matrix}[i-1][j-1]))^2\}^{\frac{1}{2}};$
8. return $C_{matrix}[x_{len}][y_{len}];$

Fig: 1 DTW algorithm

4. **Wavelet based Featured Time Warping**

In this paper, we propose two new methods for retrieval of similar sequences from large databases. Our first method uses the features extracted from wavelet based compressed signal and second method uses the featured vector of decomposed coefficients, both support indexed based time warping distance measure. Our main objective is to improve the search performance in huge databases without allowing the false dismissals and minimization of false alarms. To achieve this goal, we introduce wavelet based featured time warping distance functions $D_{tw-wlt}comp$ and $D_{tw-wlt}coeff$ for our proposed methods, both under estimate and lower bound the original time warping distance efficiently and consistently. Our distance measures also satisfy the triangular inequality, a precondition for indexing in most of the metric spaces [1][23]. In preprocessing phase, first of all we decompose the network traffic signals using wavelet-based techniques. In the first phase for first proposed method, we smooth the signal while using wavelet filtering techniques and then compressed them by retaining 99% energy of original signal using global threshold methods and select the features of the compressed signal on the basis of its maxima, minima and average. In the second proposed method, we decompose the signal and select the featured vector from the coefficients of decomposed signal using the average of approximation coefficients at the coarsest scale, maxima of maxima and minima of minima of detail coefficients.

For the first method, we can write the Eq. (6) as

$$f(t) \cong S_J(t) + \sum_{i=1}^{J} D_i(t) \qquad (7)$$

which can be further reduced using global threshold methods to

$$f(t) \cong S_J(t) + \sum_{i=1}^{p \le J} D_i(t) \qquad (8)$$

The Eq. (8) will be our reduced model on the basis of which we will compress our signal. The compression rate will be the ratio between the length of reduced model and length of original signal. If we write the synthesized compressed signal as

$$Comp\ f(t) \cong \{a_1, a_2, \cdots, a_n\} \qquad (9)$$

then for efficient access to similar sequences in large databases, we select a 3-tuple feature vector out of each compressed signal on the basis of their maxima, minima and average denoted by $(a_{\max})$, $(a_{\min})$ and (a_{ave}) respectively. We introduce this feature vector as indexing attribute to multi-dimensional index and $D_{tw-wlt}comp$ as distance function to scan the similar sequences in pre-filtering phase using range query. In our second proposed method, the model of decomposition given in Eq. (7) will be truncated to

$$f(t) \cong S_J(t) + \sum_{i=1}^{J} \max D_i(t) + \sum_{i=1}^{J} \min D_i(t) \qquad (10)$$

We select 3-tuple feature vector from this model as indexing attribute on the basis of following:

a. The mean of the approximation signal at the coarsest scale, i.e. $\bar{a}_J$
b. The maxima of maxima of detail coefficient at all scales, i.e. $\max d_{\max}$
c. The minima of minima of detail coefficient at all scales, i.e. $\min d_{\min}$

and $D_{tw-wlt}coeff$ as distance metric. In post processing step, for both these methods we again compute the time warping distance of original signals for truncated database to avoid any false alarms. This will increase our execution time but gives more accuracy. Note that in our proposed algorithms the speed comes from less distance calculations for featured vectors in first phase and distance calculations for reduced database in the second phase without much compromising on accuracy. To the best of our knowledge, our approaches using the features extracted from compressed wavelet based signals and coefficients of decomposed signals on the basis of averages, maxima and minima are not explored in the related work for mining similarities in large data basis.

Our Proposed methods are given in algorithm 2.

Algorithm 2: Wavelet based featured time warping
Input: S, Q Output: = O
1. Decomposition and compression of query Q and database $S = \{s_1, s_2, \cdots, s_i \cdots s_N\}$ where $i = 1,2,\cdots N$ using wavelets. Each s_i is having arbitrary length either $\{2^{16}, 2^{18}, 2^{20}, 2^{22}\}$.
2. Extraction of featured vectors of Q and S from compressed signal at low resolution on the basis of their maxima, minima and average.
3. Given $F(Q)$ and $F(S_i)$ in step 2, perform range query on multi-dimensional index using R-tree. If
(a) for method 1 $D_{tw-wlt}comp(F(Q), F(S_i)) \leq \varepsilon$ and
(b) for method 2 $D_{tw-wlt}coeff(F(Q), F(S_i)) \leq \varepsilon$
then add to output O
4. For each i in answer if $D_{tw}comp(Q, S_i) > \varepsilon$ *or* $D_{tw}coef(Q, S_i) > \varepsilon$ then remove i from O
5. Return O

4.1 SSTDB and Analysis of Proposed Algorithm

We have applied our experiments on a large database named SSTDB (Self similar network traffic database). This database contains 7150 total network traffic signals (550 files in each of 13 folders). All of these files were than be packed in *.rar format using winRar software of 300MB packages and uploaded on [24]. The total size of this database is approximately 30GB.

We selected the length of sequences ranging 2^{16} to 2^{22} (with even powers only) observations at a dyadic scale. The queries were also selected from same database randomly. We reported the results are much better for huge data bases [1][12]. For all the experiments, we developed our own codes for wavelet transformations and multidimensional index search within the MATLAB environment.

4.2 Comparisons and Evaluation of Results

We compare our proposed methods with the original time warping distance metric D_{tw} in which we select a sequence from a database and compute its time warping distance with a given query sequence using dynamic time warping algorithm to search their similarity and repeat the process for all the sequences in the database. We also compare our results with so far claimed the best lower bounding distance function D_{tw-lb} [22]. We use the L^2 norm for base distance function for all these methods. We carried out our experiments using Haar, db4, and sym6 wavelets and reported the results for Haar wavelet due to its simplicity. We used the range query and also employed the multi-dimensional index strategy using R-tree upon featured vector spaces of our proposed methods. In all the experiments, we fixed the values of query thresholds and values of precision and recall, two negatively correlated quantities, for the sake of comparison of these techniques. All our results are averaged over 50 trials for each experiment.

4.2.1 Experiment 1 Varying the number of sequences

Our first experiment evaluates the performance efficiency of our proposed methods in terms of their elapsed time with the data set and query sequences selected from the randomly generated synthetic data. The CPU time measurement starts when the query is posed and ends when the answers are retained after the post-processing step.

In this experiment we used varied the number of sequences of arbitrary lengths ranging from 2^{16} to 2^{22} (with even powers only) observations for data base size of 100 to 500 sequences and reported the results averaged over 50 trials. Figure 2 and table 1 show the average elapsed CPU time for all four methods. Our proposed methods give much better results as compared to original time warping distance D_{tw} and compatible results with so far claimed best D_{tw-lb} technique. Our second proposed method based on the distance metric $D_{tw-wlt}coeff$ has given much better results as compared to all three methods and almost constant CPU time for even large data bases.

Cat.	Original	D_{tw-lb}	$D_{tw-wlt}comp$	$D_{tw-wlt}coeff$
#of Seq	Average CP Time	Average CP Time	Average CP Time	Average CP Time
100	1092	827	774	751
200	1611	1036	921	750
300	2022	1212	1046	751
400	2467	1395	1175	755
500	2954	1595	1317	756

Table 1: Results for varying number of sequences of arbitrary length (average CPU time in seconds)

4.2.3 Experiment 2: Varying the length of sequences

This experiment is again on synthetic data in which we have varied the length of sequences to 2^{16} to 2^{22} (with even powers only).We fixed the data base size to 100 sequences and reports the results for average of 50 trials for each sequence length. Figure 3 and table 2 shows the average elapsed CPU time for the four methods. Our proposed methods give four to eleven times better results depending on the number of sequences as compared to D_{tw-lb} and almost twenty times to original time warping distance D_{tw}.

The distance metric selected on the basis of wavelet coefficients $D_{tw-wlt}coeff$ gives even better results as compared to $D_{tw-wlt}comp$ in most of the cases. The results also show that the CPU time in both of our proposed cases do not increase rapidly with the increase in number of sequences as compared to D_{tw} and D_{tw-lb}.

Cat	Original	D_{tw-lb}	$D_{tw-wlt}comp$	$D_{tw-wlt}coeff$
Lof Seq	Average CP Time	Average CP Time	Average CP Time	Average CP Time
2^{16}	646	636	633	632
2^{18}	715	667	635	634
2^{20}	1434	904	659	664
2^{22}	6478	2617	926	897

Table 2: Results of varying number of sequences for fixed length (average CPU time in seconds)

4.2.4 Experiment 3: Varying the number of sequences for fixed length

Our next experiment is again on synthetic data in which we varied the number of sequences from 500 to 2000 with a gap of 500 for length of 2^{18} with the same experimental settings discussed in previous experiments. Figure 4 and table 3 shows the average elapsed CPU time for the four methods. It is observed that all four methods give linearly increased CPU time with the increasing number of sequences. Our second proposed method $D_{tw-wlt}coeff$ gives almost 15 to 70 times better results as compared to original time warping distance D_{tw} and 2 to 10 times to D_{tw-lb} .However in this experiment D_{tw-lb} gives slightly better results as compared to our first proposed method $D_{tw-wlt}comp$. Figure 5 shows the graph of similar sequences extracted from the large database by using our proposed algorithms.

Cat	Original	D_{tw-lb}	$D_{tw-wlt}comp$	$D_{tw-wlt}coeff$
# of Seq	Average CP Time	Average CP Time	Average CP Time	Average CP Time
500	653	635	638	635
1000	672	639	644	638
1500	692	643	650	641
2000	711	646	656	643

Table 3: Results for varying number of sequences of fixed length (average CPU time in seconds)

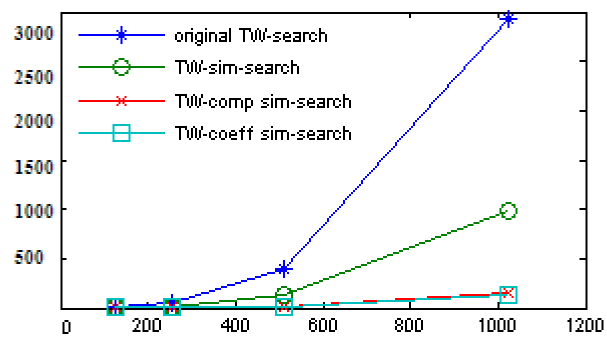


Fig. 2 Comparison of average elapsed CPU time by varying the number of sequences

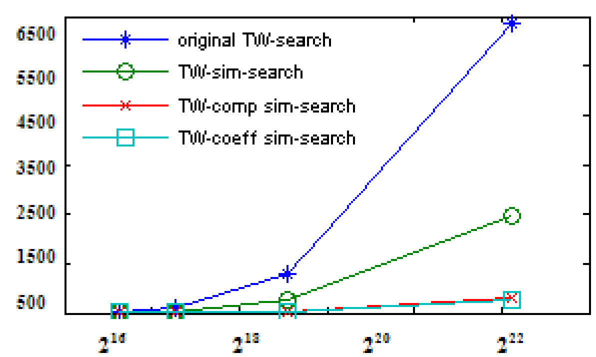


Fig. 3: Comparison of average elapsed CPU time by varying the length of sequences.

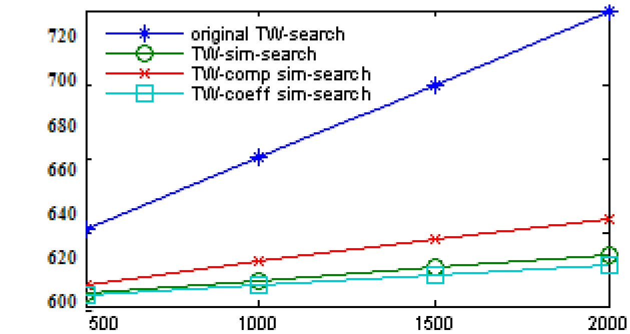


Fig. 4: Comparison of elapsed CPU time by varying the number of sequences of fixed length.

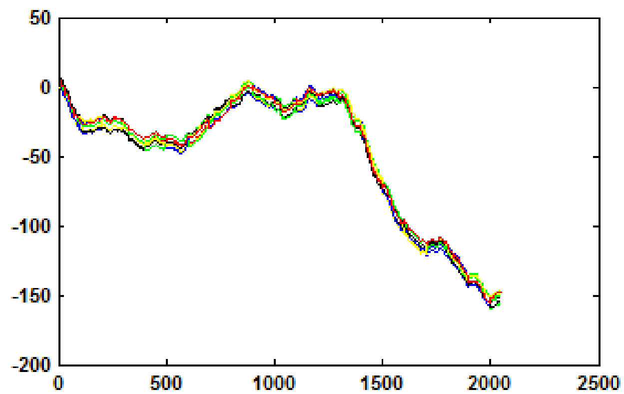


Fig. 5: Graph of similar collected sequences using proposed algorithm.

5 Discussion and Conclusions

In this paper we present two new approaches to wavelet transformation for retrieval of similar sequences using indexed based time warping technique. We proposed two lower bounding distance functions $D_{tw-wlt}comp$ and $D_{tw-wlt}coeff$ using 3-tuple feature vector on the basis of maxima, minima and average extracted from the wavelet based compressed signal for the first proposed method and 3-tuple feature vector from decomposed wavelet coefficients on the basis of over all average of approximation coefficients at the coarsest level and maxima of maxima and minima of minima of detail coefficients for our second proposed model. Our proposed methods are based on the theory that most of the time series signals are always having the noise and unnecessary details not required for further analysis hence features extracted directly from such signals may not give the true picture of it. A compressed signal using wavelet transformation filters will be comparatively more smoothed and features extracted from it are more helpful for speedy and accurate search procedures as proved in experiments conducted on the basis of our first proposed method. At the same time, in our second proposed method, features extracted from the wavelet decomposition coefficients capture the over all trend and the important discontinuities of the signal which are severely smoothed out by the compression process in the first method. We proved that our proposed methods are compatible with prevailing methods and even much better in most of the cases particularly for our second proposed method based on features extracted from wavelet coefficients and distance metric $D_{tw-wlt}coeff$ .We carried out experiments with different wavelet families and found our methods applicable to all these. We have left the comparison of all these for some future work and reported the results with the simplest, cost effective and efficiently calculated Haar wavelets. In future we intend to explore vagueness of similarity in large time series databases based on our proposed techniques.

References

[1] R. Agrawal, C.Faloutsos, and A.Swami. Efficient similarity search in sequence databases. In Proc. of the FODO Conf., volume 730. Springer, 1993.

[2] K.V.R Kanth, D.Agrawal, and A.K. Singh.Dimensionality reduction for similarity searching in dynamic databases. In Proc. of the ACM SIGMOD Conf., pages 166-176, seattle, Washington, USA, 1998.

[3] E. Keogh, K. Chakrabarti, M. Pazzani, and S.Mehrotra. Dimensionality reduction for fast similarity search in large time series databases. In Knowledge and Information Systems, volume 3, pages 263-286, 2001.

[4] D. Rafiei and A. Mendelzon. Similarity based queries for time series data. In Proc. of the ACM SIGMOD Conf., Tucson, Arizona, USA, May, 1997.

[5] B.K. Yi and C. Faloutsos. Fast time sequence indexing for arbitrary L_p norms. In Proc. of the VLDB Conf., pages 385-394, Cairo, Egypt, 2000. Morgan Kaufmann.

[6] D.Wu, D.Agrawal, A.E. Abbadi, A.K. Singh.and T.R. Smith. Efficient retrieval of browsing large image databases. In Proc. of CIKM '96, pages 11-18, Rockville, Maryland, USA, November 1996.

[7] C. Faloutsos, M. Ranganathan, and Y. Manolopoulos. Fast subsequence matching in time series databases.In Proc. of the ACM SIGMOD Conf., pages 419-429, Minneapolis, Minnesota, USA, 1994.

[8] D. Rafiei and A. Mendelzon. Efficient retrieval 0f Similarity time sequences using DFT. In Proc. of the FODO Conf., Kobe, Japan, November, 1998.

[9] K. Chan and A.W. C. Fu. Efficient time series matching by wavelets. In Proc. of the ACDE Conf. pages 126-133, Sydney, Australia, 1999.

[10] Y.L. Wu, D. Agrawal, A.E Abbadi. A comparison of DFT and DWT based similarity search in time series databases. In Proc. of the F5th Conf. on Knowledge Information, pages 11-18, 2000.

[11] Percival, D.B., Walden, A.T., Wavelet Methods for Time Series Analysis,Cambridge University Press,2002.

[12] Daubechies, I., Ten Lectures on Wavelets. Society for Industrial and Applied Mathematics. Philadelphia, Pennsylvania,1992.

[13] Cohen A., Numerical analysis of wavelet methods, Studies in Mathematics and Its Applications, vol.32, 2003, Elsevier publishers.

[14] Mallat S. G., A theory of multiresolution signal decomposition: the wavelet representation. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 11, no. 7, pp. 674-693,1989.

[15] Keogh, E., Pazzani, M. Scalling up dynamic time warping for data mining application. In 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Boston, 2000.

[16] Yi, B., Jagadish, H and Faloutsos,C. Efficient retrieval of similar time sequences under time warping. In International Conference of Data Engineering, 1998.

[17] Berndt,D., Clifford, J. Using dynamic time warping to find patterns in time series. AAAI-94 Workshop on Knowledge Discovery in Databases (KDD-94), Seattle, Washington, 1994.

[18] Gavrila, D. M.,Davis, L.S. Towards 3-d model based tracking and recognition of human movement: a multi-view approach. In International Workshop on Automatic Face and Gesture Recognition. IEEE Computer Society, Zurich, 1995.

[19] Schmill, M., Oates, T., Cohen, P. Learned models for continuous planning. In Seventh International Conference on Artificial Intelligence and Statistics, 1999.

[20] Gollmer, K., Posten, C. Detection of distorted pattern using dynamic time warping algorithm and application for supervision of bioprocesses. On Line Fault Detection and Supervision in the Chemical Process Industries, 1995.

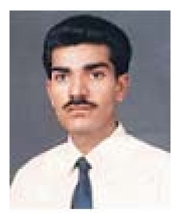
[21] Caiani, E.G., Porta, A., Baselli, G., Turiel, M., Muzzupappa, S., Pieruzzi, F., Crema, C., Malliani, A., Cerutti, S. Warped Average template technique to track on a cycle by cycle basis to cardiac filling phases on left ventricular volume. IEEE computers in Cardiology ,vol 25, NY,USA, 1998.

[22] Kim S, Park S, Chu W. An index- based approach for similarity search supporting time warping in large sequence databases. In Proceedings of the 17th International Conference on Data Engineering, page 607-614, 2001.

[23] J. Foran, Fundamentals of real analysis. Marcel Deckker, 1991.

[24] Self-Similar Traffic Database (2010) http://www.filefactory.com/f/fad37c19c06c22b9/

Dr. S. M. Aqil Burney is the Director & Chairman at Department of Computer Science, (UBIT), University of Karachi. His research interest includes AI, Soft Computing, Neural Networks, Fuzzy Logic, Data Mining, Statistics, Simulation and Stochastic Modeling of Mobile Communication System and Networks and Network Security, Computational Biology and Bioinformatics. He is a senior member of IACSIT



Akhter Raza Syed is a research fellow at University of Karachi, Pakistan. Received his M.Sc. Degree from University of Karachi in 1991. The M.Phil. work converted to Ph.D. in 2006 and leading towards Ph.D. Working as Assistant Professor in Education Department. Worked as visiting faculty member in Institute of Business Administration (**IBA Karachi**) from 2004 to 2009. Also teaching as visiting faculty in Umaer Basha Institute of Technology (UBIT) University of Karachi, Pakistan. He is a senior member of IACSIT

Lt. Col. Dr. Afzal Saleemi got his bachelor's degree in 1986 and master's degree in 1989 from the University of Punjab. Completed his Ph.D. in computer science in 2007, Has remained on various instructional as well as administrative appointments and is having more than 20 years experience of teaching. Presently, serving as Head of Computer Science Department National University of Sciences and Technology, NUST (PMA).