A Dimension-Reduction Framework for Human Behavioral Time Series Data

Santi Phithakkitnukoon and Ram Dantu
Department of Computer Science & Engineering
University of North Texas
Denton, TX, 76203 USA
{santi | rdantu}@unt.edu

Abstract

Human-machine interaction has become one of the most active research areas, and influenced several new paradigms of computing such as Social computing, Mobile computing, and Pervasive/Ubiquitous computing, which are typically concerned with the study of human user's behavior to facilitate behavioral modeling and prediction. Human behavioral data are usually high-dimensional time series, which need dimension-reduction strategies to improve the efficiency of computation and indexing. In this paper, we present a dimension-reduction framework for human behavioral time series. Generally, recent behavioral data are much more interesting and significant in understanding and predicting human behavior than old ones. Our basic idea is to reduce to data dimensionality by keeping more detail on recent behavioral data and less detail on older data. We distinguish our work from other recent-biased dimension-reduction techniques by emphasizing on recent-behavioral data and not just recent data. We experimentally evaluate our approach with synthetic data as well as real data. Experimental results show that our approach is accurate and effective as it outperforms other well-known techniques.

Introduction

Time series is a sequence of time-stamped data points. It is used to represent different types of data such as stock price, exchange rate, temperature, humidity, power consumption, and event logs. Time series are typically large and of high dimensionality, which introduces the "curse of dimensionality" problem in machine learning. To improve the efficiency of computation and indexing, dimension-reduction techniques are needed for high-dimensional data. Among the most widely used techniques are PCA (Principal Component Analysis) (also known as Singular Value Decomposition (SVD)), Discrete Wavelet Transform (DWT), and Discrete Fourier Transform (DFT). Other recently proposed techniques are PIP (Perpetually Important Points) (Fu *et al.* 2001), PAA (Piecewise Aggregate Approximation) (Keogh *et al.* 2000), and landmarks (Perng *et al.* 2000). These techniques were developed to reduce the dimensionality of the time series by considering every part of a time series equally. In many applications such as the stock market, however, recent data are much more interesting and significant than old data, "recent-biased analysis" (the term originally coined by Zhao and Zhang 2006) thus emerges. The recently proposed techniques include SWAT (Bulut and Singh 2003), equi-DFT (Zhao and Zhang 2006), vari-DFT (Zhao and Zhang 2006), and others (Aggarwal *et al.* 2003; Chen *et al.* 2002; Giannella *et al.* 2003; Palpanas *et al.* 2004).

With Ambient Intelligence (AmI) (Remagnino and Foresti 2005) and new paradigms of computing, *e.g.,* Social computing, Mobile computing, and Pervasive/Ubiquitous computing, more human behavioral time series are being processed and analyzed to model human-machine interaction. Human behavior data are being collected from many sources such as sensors, mobile devices, and wearable computers. Typically, human behavior tends to repeat periodically, which creates a pattern that changes over different periods. This change of behavioral pattern distinguishes human behavioral time series from many other types of time series as well as provides the key to our proposed framework in dimension reduction particularly for human behavioral time series data. Since human behavioral pattern changes over time, the most recent pattern is more significant than older ones. In this paper, we introduce a new recent-pattern biased dimension reduction framework that gives more significance to the recent-pattern data (not just recent data) by keeping it with finer resolution, while older data is kept at coarser resolution. We distinguish this paper from other previously proposed recent-biased dimension-reduction techniques by the following contributions:

1. We introduce a new framework for dimension reduction for human behavioral time series by keeping more detail on data that contains the most recent pattern and less detail on older data.
2. Within this framework, we also propose Hellinger distance-based algorithms for recent periodicity detection and recent-pattern interval detection.

Dimension Reduction for Human Behavioral Time Series

In dynamic stream data analysis, changes in recent data usually receive more attention than old data. Human behavioral time series is also stream data at a variety rate. This rate of data generation depends on the type of behaviors that are being monitored. Nevertheless, changes in the recent data are normally more significant than the old data. Especially, human behavior tends to repeat periodically, which creates a pattern that alters over many periods due to countless factors. Generally, future behavior is more relevant to the recent behavior than the older ones. Our main goal in this work is to reduce dimensionality of a time series generated from human behavior. The basic idea is to keep data that contains recent pattern with high precision and older data with low precision. Since the change in human behavior over time creates changes in the pattern and the periodicity rate, thus we need to detect the most recent periodicity rate which will lead to identifying the most recent pattern. Hence a dimension reduction technique can then be applied. This section presents our novel framework for dimension reduction for human behavioral time series data, which includes new algorithms for recent periodicity detection, recent-pattern interval detection, and dimension reduction.

Recent Periodicity Detection

Unlike other periodicity detection techniques (Berberidis *et al.* 2002; Elfeky, Aref, and Elmagarmid 2004; Elfeky, Aref, and Elmagarmid 2005; Indyk, Koudas, and Mathukrishnan 2000; Ma and Hellerstein 2001; Yang, Wang, and Yu 2000) that attempt to detect the global periodicity rates, our focus here is to find the "most recent" periodicity rate of time series data. Let X denote a time series with N time-stamped data points, and x_i be the value of the data at time-stamp i. The time series X can be represented as $X = x_0, x_1, x_2, ..., x_N$, where x_0 is the value of the most recent data point and x_N is the value of the oldest data point. Let $\Phi(k)$ denote the recent-pattern periodicity likelihood (given by (1)) that measures the likelihood of selected recent time segment (k) being the recent period of the time series, given that the time series X can be sliced into equal-length segments $X_0^k, X_1^k, X_2^k, ..., X_{\lfloor N/k \rfloor -1}^k$, each of length k, where $X_i^k = x_{ik}, x_{ik+1}, x_{ik+2}, ..., x_{ik+k-1}$.

$$\Phi(k) = \frac{\sum_{i=1}^{\lfloor N/k \rfloor -1} (1 - d_H^2(\hat{X}_0^k, \hat{X}_i^k))}{\lfloor N/k \rfloor - 1}, \qquad (1)$$

where $d_H^2(A, B)$ is Hellinger distance (Yang and Cam 2000), which is widely used for estimating a distance (difference) between two probability measures (*e.g.*, probability density functions (pdf), probability mass functions (pmf)). Hellinger distance between two probability measures A and B can be computed by (2). A and B are M-tuple $\{a_1, a_2, a_3, ..., a_M\}$ and $\{b_1, b_2, b_3, ..., b_M\}$ respectively, and satisfy $a_m \geq 0, \sum_m a_m = 1, b_m \geq 0$, and $\sum_m b_m = 1$. Hellinger distance of 0 implies that $A = B$ whereas disjoint A and B yields the maximum distance of 1.

$$d_H^2(A, B) = \frac{1}{2} \sum_{m=1}^{M} (\sqrt{a_m} - \sqrt{b_m})^2. \qquad (2)$$

In our case, $\hat{X}_0^k$ and $\hat{X}_i^k$ are X_0^k and X_i^k after normalization, respectively, such that they satisfy the above conditions. Thus, $\Phi(k)$ has the values in the range [0, 1] as 0 and 1 imply the lowest and the highest recent-pattern periodicity likelihood, respectively.

Definition 1 If a time series X of length N can be sliced into equal-length segments $X_0^p, X_1^p, X_2^p, ..., X_{\lfloor N/p \rfloor -1}^p$, each of length p, where $X_i^p = x_{ip}, x_{ip+1}, x_{ip+2}, ..., x_{ip+p-1}$, and $p = \arg\max_k \Phi(k)$, then p is said to be the recent periodicity rate of X.

The basic idea of this algorithm is to find the time segment (k) that has the maximum $\Phi(k)$, where $k = 2, 3, ..., \lfloor N/2 \rfloor$. If there is a tie, smaller k is chosen to favor shorter periodicity rates, which are more accurate than longer ones since they are more informative (Elfeky *et al.* 2005). The detailed algorithm is given in Algorithm 1. Note that $\Phi(1) = 1$ since $d_H^2(\hat{X}_0^1, \hat{X}_i^1) = 0$, hence k begins at 2.

Algorithm 1 Recent Periodicity Detection
p = PERIODICITY(X)
Input: Time series (X) of length N
Output: Recent periodicity rate (p)
1. FOR k = 2 to $\lfloor N/2 \rfloor$
2. Compute $\Phi(k)$;
3. END FOR
4. $p = k$ that maximizes $\Phi(k)$;
5. IF $|k| > 1$
6. $p = min(k)$;
7. END IF
8. Return p as the recent periodicity rate;

Recent-Pattern Interval Detection

After obtaining the recent periodicity rate p, our next step towards dimension reduction for a time series X is to detect the time interval that contains the most recent pattern. This interval is a multiple of p. We base our detection on the *shape* of the pattern and the *amplitude* of the pattern.

For the detection based on the shape of the pattern, we construct three Hellinger distance-based matrices to measure the differences within the time series as follows:

1. $D_1^i = [\ d_1(1) \quad d_1(2) \quad ... \quad d_1(i)\ ]$ is the matrix whose elements are Hellinger distances between the pattern derived from the X_0^p to X_{j-1}^p ($\bar{X}_{0 \to j-1}^p$), which can be simply computed as a mean time series over time segments 0 to $j-1$ given by (4), and the pattern captured within the time segment j (X_j^p) as follows:

$$d_1(j) = d_H^2(\hat{\bar{X}}_{0 \to j-1}^p, \hat{X}_j^p), \qquad (3)$$

where

$$\bar{X}_{0 \to j-1}^p = \frac{1}{j} \sum_{n=0}^{j-1} x_{np}, \frac{1}{j} \sum_{n=0}^{j-1} x_{np+1}, ..., \frac{1}{j} \sum_{n=0}^{j-1} x_{np+p-1}. \qquad (4)$$

Again, the hat on top of the variable indicates the normalized version of the variable.

2. $D_2^i = [\ d_2(1) \quad d_2(2) \quad ... \quad d_2(i)\ ]$ is the matrix whose elements are Hellinger distance between the most recent

pattern captured in the first time segment (X_0^p) and the pattern occupied within the time segment j (X_j^p) as follows:

$$d_2(j) = d_H^2(\hat{X}_0^p, \hat{X}_j^p). \quad (5)$$

3. $D_3^i = [\ d_3(1) \quad d_3(2) \quad \ldots \quad d_3(i)\ ]$ is the matrix whose elements are Hellinger distance between the adjacent time segments as follows:

$$d_3(j) = d_H^2(\hat{X}_{j-1}^p, \hat{X}_j^p). \quad (6)$$

These three matrices provide the information on how much the behavior of the time series changes across all time segments. The matrix D_1^i collects the degree of difference that X_j^p introduces to the recent segment(s) of the time series up to $j = i$, where $j = 1, 2, 3, ..., \lfloor N/p \rfloor - 1$. The matrix D_2^i records the amount of difference that the pattern occupied in the time segmentX_j^p makes to the most recent pattern captured in the first time segmentX_0^p up to $j = i$. The matrix D_3^i keeps track of the differences between the patterns captured in the adjacent time segmentsX_{j-1}^p and X_j^pup to $j = i$.

To identify the recent-pattern interval based on the shape of the pattern, the basic idea here is to detect the first change of the pattern that occurs in the time series as we search across all the time segments X_j^p in an increasing order of j starting from $j = 1$ to $\lfloor N/p \rfloor - 1$. Several changes might have been detected as we search through entire time series, however our focus is to detect the most recent pattern. Therefore, if the first change is detected, the search is over. The change of pattern can be observed from the significant changes of these three matrices. The significant change is defined as follows.

Definition 2 If $\mu_{D_k^i}$ and $\sigma_{D_k^i}$ is the mean and the standard deviation of D_k^i and $\mu_{D_k^i} + 2\sigma_{D_k^i} \leq d_k(i+1)$, then X_{i+1}^p is said to make the significant change based on its shape.

Algorithm 2 Significant Change Detection
y = SIG_CHANGE(D_k^i,$d_k(i+1)$)
Input: Distance matrix (D_k^i) and the corresponding distance element$d_k(i+1)$.
Output: Binary output (y) of 1 implies that there is a significant change made by X_{i+1}^p and 0 implies otherwise.
1. IF $\mu_{D_k^i} + 2\sigma_{D_k^i} \leq d_k(i+1)$
2. y = 1;
3. ELSE
4. y = 0;
5. END IF

With the detected significant changes in these distance matrices, the recent-pattern interval based on the shape of the pattern can be defined as follows. The detailed algorithm is given in Algorithm 3.

Definition 3 If X_{i+1}^p introduces a significant change to at least two out of three matrices (D_1^i,D_2^i, and D_3^i), then the recent-pattern interval based on the shape (r_{shape}) is said to be ip time units.

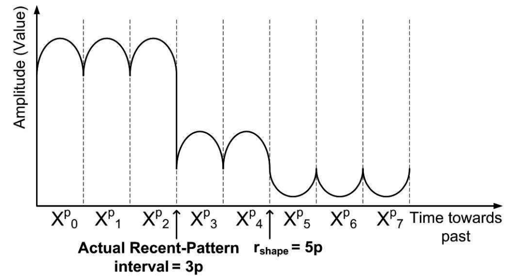

Figure 1: An example of misdetection for the recent-pattern interval based on the shape of the pattern. Algorithm 3 would detect the change of the pattern at the 5^{th} time segment (X_5^p) whereas the actual significant change takes place at the 3^{rd} time segment (X_3^p).

Algorithm 3 Shape-based Recent-Pattern Interval Detection
r_{shape} = SHAPE_RPI($D_1^{\lfloor N/p \rfloor-1}$,$D_2^{\lfloor N/p \rfloor-1}$,$D_3^{\lfloor N/p \rfloor-1}$)
Input: Three distance matrices ($D_1^{\lfloor N/p \rfloor-1}$,$D_2^{\lfloor N/p \rfloor-1}$, $D_3^{\lfloor N/p \rfloor-1}$).
Output: Shape-based recent-pattern interval (r_{shape}).
1. Initialize r_{shape} to N
2. FOR i = 2 to $\lfloor N/p \rfloor - 1$
3. IF SIG_CHANGE(D_1^i,$d_1(i$+1)) + SIG_CHANGE(D_2^i,$d_2(i$+1)) + SIG_CHANGE(D_3^i,$d_3(i$+1)) = 2
4. $r_{shape} = ip$;
5. EXIT FOR LOOP
6. END IF
7. END FOR
8. Return r_{shape} as the recent-pattern interval based on the shape;

For this shape-based recent-pattern interval detection, the Hellinger distances are computed by taking the normalized version of the patterns in the time segments. Since normalization rescales the amplitude of the patterns, the patterns with similar shapes but significantly different amplitudes will not be detected (see an example illustrated in Figure 1).

To handle this shortcoming, we propose an algorithm to detect the recent-pattern interval based on the amplitude of the pattern. The basic idea is to detect the significant change in the amplitude across all time segments. To achieve this goal, let $A^i = [\ a(1) \quad a(2) \quad \ldots \quad a(i)\ ]$ denote a matrix whose elements are mean amplitudes of the patterns of each time segment up to time segment i, which can be easily computed by (7).

$$a(k) = \frac{1}{p} \sum_{n=0}^{p-1} x_{(k-1)p+n}. \quad (7)$$

Similar to the previous case of distance matrices, the significant change in this amplitude matrix can be defined as follows.

Definition 4 If μ_{A^i} and σ_{A^i} is the mean and the standard deviation of A^i and $\mu_{A^i} + 2\sigma_{A^i} \leq a(i+1)$, then X_{i+1}^p is said to make the significant change based on its amplitude.

Likewise, with the detected significant change in the amplitude matrix, the recent-pattern interval based on the amplitude of the pattern can be defined as follows. The detailed algorithm is given in Algorithm 4.

Definition 5 If X_{i+1}^p makes a significant change in the matrix (A^i), then the recent-pattern interval based on the amplitude (r_{amp}) is said to be ip time units.

Algorithm 4 Amplitude-based Recent-Pattern Interval Detection
r_{amp} = AMP_RPI($A^{\lfloor N/p \rfloor - 1}$)
Input: The amplitude matrix ($A^{\lfloor N/p \rfloor - 1}$).
Output: Amplitude-based recent-pattern interval (r_{amp}).
1. Initialize r_{amp} to N
2. FOR i = 2 to $\lfloor N/p \rfloor - 1$
3. IF SIG_CHANGE(A^i,$a(i+1)$) = 1
4. r_{amp} = ip;
5. EXIT FOR LOOP
6. END IF
7. END FOR
8. Return r_{amp} as the recent-pattern interval based on the amplitude;

Finally, the recent-pattern interval can be detected by considering both shape and amplitude of the pattern. Based on the above algorithms for detecting the interval of the most recent pattern based on the shape and the amplitude of the pattern, the final recent-pattern interval can be defined as follows.

Definition 6 If r_{shape} is the recent-pattern interval based on the shape of the pattern and r_{amp} is the recent-pattern interval based on the amplitude of the pattern, then the final recent-pattern interval(R) is the lowest value among r_{shape} and r_{amp} – *i.e.*, $R = min(r_{shape}, r_{amp})$.

Dimension Reduction

Our main goal in this work is to reduce dimensionality of a human behavioral time series. The basic idea is to keep more details for recent-pattern data, while older data kept at coarser level.

Based on the above idea, we propose a dimension reduction scheme for human behavior time series data that applies a dimension reduction technique to each time segment and then keeps more coefficients for data that carries recent-behavior pattern and fewer coefficients for older data.

Several dimension reduction techniques can be used in our framework such as DWT, DFT, SVD, and others. In this paper, we choose DFT for demonstration.

Let C_i represent the number of coefficients retained for the time segment X_i^p. Since our goal is to keep more coefficients for the recent-pattern data and fewer coefficients for older data, a sigmoid function (given by (8)) is generated and centered at R time units (where the change of behavior takes place).

$$f(t) = \frac{1}{1 + \alpha^{-t/p}}. \tag{8}$$

The decay factor (α) is automatically tuned to change adaptively with the recent-pattern interval (R) by being set

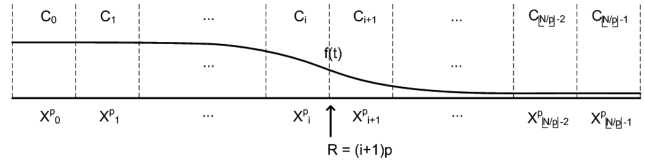

Figure 2: Dimension reduction scheme for human behavioral time series. A time series is partitioned into equal-length segments of length p (recent periodicity rate) and more coefficients are taken for recent-pattern data and fewer coefficients are taken for older data based on the decay rate of a sigmoid function ($f(t)$). For this example, recent-pattern interval (R) is assumed to be $(i+1)p$.

at $\alpha = p/R$, such that a slower decay rate is applied to a longer R and vice versa. The number of coefficients for each time segment can be computed as the area under the sigmoid function over each time segment (given by (9)), so the value of C_i is within the range [1, p].

$$C_i = \left\lceil \int_{X_i^p} f(t)dt \right\rceil. \tag{9}$$

C_i decreases according to the area under the sigmoid function across each time segment as i increases, hence $C_0 \geq C_1 \geq C_2 \geq ... \geq C_{\lfloor N/p \rfloor - 1}$. For each time segment, we choose the first C_i coefficients that capture the low-frequency part of the time series.

With this scheme, a human behavioral time series data can be reduced by keeping the more important portion of data (recent-pattern data) with high precision and the less important data (old data) with low precision. As future behavior is generally more relevant to the recent behavior than old ones, maintaining the old data at low detail levels might as well reduces the noise of the data, which would benefit predictive modeling for (individual and group) human behavior. This scheme is shown in Figure 2, and the detailed algorithm is given in Algorithm 5.

Note that if no significant change of pattern is found in the time series, our proposed framework will work similarly to equi-DFT as our R is initially set to N (by default, see Algorithm 3, Algorithm 4, and Definition 6). Hence the entire series is treated as a recent-pattern data, *i.e.*, more coefficients are kept for recent data and fewer for older data according to (the left-hand side from the center of) the sigmoid function with decay factor $\alpha = p/R$.

Algorithm 5 Dimension Reduction for Human Behavioral Time Series
Z = DIMENSION_REDUCTION(X)
Input: A human behavioral time series (X) of length N.
Output: A reduced time series (Z).
1. p = PERIODICITY(X);
2. Partition X into equal-length segments, each of length p;
3. Compute matrices $D_1^{\lfloor N/p \rfloor - 1}$, $D_2^{\lfloor N/p \rfloor - 1}$, $D_3^{\lfloor N/p \rfloor - 1}$, and $A^{\lfloor N/p \rfloor - 1}$;
4. r_{shape} = SHAPE_RPI($D_1^{\lfloor N/p \rfloor - 1}$,$D_2^{\lfloor N/p \rfloor - 1}$,$D_3^{\lfloor N/p \rfloor - 1}$);

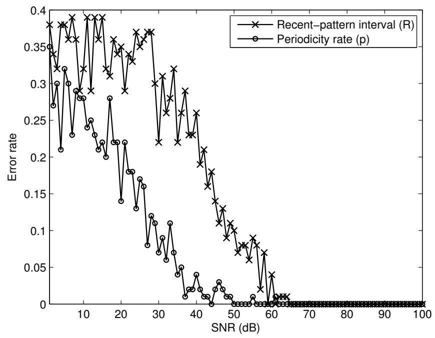

Figure 3: Experimental result of the error rate at differe SNR levels of 100 synthetic time series (with known p ar R).

5. r_{amp} = AMP_RPI($A^{\lfloor N/p \rfloor -1}$);
6. $R = min(r_{shape}, r_{amp})$;
7. Place a sigmoid function $f(t)$ at R;
8. FOR each segment i
9. $Coefs\_DFT$ = apply DFT for segment i;
10. Compute C_i;
11. z_i = C_i first $Coefs\_DFT$;
12. END FOR
13. $Z = \{z_0, z_1, z_2, ..., z_{\lfloor N/p \rfloor -1}\}$; /* Series of selected coefficients */
14. Return Z as the reduced time series;

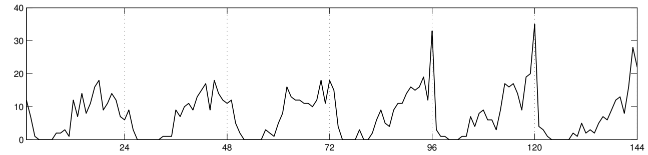

Figure 4: A monthly mobile phone usage over six months represents an individual behavioral time series with detected $p = 24$ and $R = 3p = 72$.

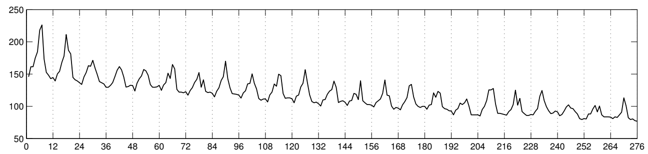

Figure 5: A monthly water usage during 1966-1988 represents a group behavioral time series with detected $p = 12$ and $R = 2p = 24$.

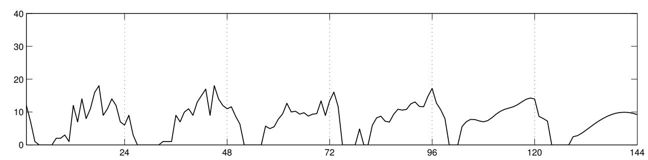

Figure 6: A reconstructed time series of the mobile phone data of 75 selected DFT coefficients from the original data of 144 data points, which is 48% reduction.

Performance Analysis

This section contains the experimental results to show the accuracy and effectiveness of our proposed algorithms. In our experiments, we exploit synthetic data as well as real data.

The synthetic data are used to inspect the accuracy of the proposed algorithms for detecting the recent periodicity rate and the recent-pattern interval. This experiment aims to estimate the ability of proposed algorithms in detecting p and R that are artificially embedded into the synthetic data at different levels of noise in the data (measured in terms of SNR (signal-to-noise ratio) in dB). For a synthetic time series with known p and R, our algorithms compute estimated periodicity rate ($\tilde{p}$) and recent-pattern interval ($\tilde{R}$) and compare with the actual p and R to see if the estimated values are matched to the actual values. We generate 100 different synthetic time series with different values of p and R. The error rate is then computed for each SNR level (0dB to 100dB) as the number of incorrect estimates (Miss) per total number of testing data, *i.e.* Miss/100. The results of this experiment are shown in Figure 3. The error rate decreases with increasing SNR as expected. Our recent periodicity detection algorithm performs with no error above 61dB while our recent-pattern interval detection algorithm performs perfectly above 64dB. Based on this experiment, our proposed algorithms are effective at SNR level above 64dB.

We implement our algorithms on two real time series data as one serves as an individual behavioral time series, and the other serves as a group behavioral time series. The first data contains the number of phone calls (both made and received) on time-of-the-day scales on a monthly basis over a period of six months (January 7^{th}, 2008 to July 6^{th}, 2008) of a mobile phone user (Phithakkitnukoon and Dantu 2008). The second data contains a series of monthly water usage (ml/day) in London, Ontario, Canada from 1966 to 1988 (Hipel and McLeod 1995). Figure 4 shows an individual behavioral time series of a mobile phone user with computed $p = 24$ and $R = 3p = 72$ based on our algorithms. Likewise, Figure 5 shows a group behavioral time series of a monthly water usage with computed $p = 12$ and $R = 2p = 24$. Based on a visual inspection, one can clearly identify that the recent periodicity rates are 24 and 12, and recent-pattern intervals are $3p$ and $2p$ for Figure 4 and 5, respectively, which shows the effectiveness of our algorithms.

We implement our recent-pattern biased dimension reduction algorithm on these two real time series data. The 144-point mobile phone data has been reduced to 75 data points (DFT coefficients), which is 48% reduction. On the other hand, the water usage data has relatively short recent-pattern interval compared to the length of the entire series thus we are able to reduce much more data. In fact, there are 276 data points of water usage data before the dimension reduction and only 46 data points are retained afterward, which is 83% reduction. The results of reconstructed time series of the mobile phone data and water usage data are shown in Figure 6 and 7, respectively.

To compare the performance of our proposed algorithm with other recent-biased dimension reduction techniques, a

Data	Percentage Reduction				Err_{RBP}				RER			
	RP-DFT	equi-DFT	vari-DFT	SWAT	RP-DFT	equi-DFT	vari-DFT	SWAT	RP-DFT	equi-DFT	vari-DFT	SWAT
Mobile phone	0.479	0.479	0.750	0.972	0.0170	0.0300	0.0330	0.191	**27.458**	15.915	22.950	5.078
Water usage	0.837	0.479	0.739	0.986	0.00712	0.00605	0.0168	0.0641	**117.550**	79.201	43.996	15.375

Table 1: Performance comparison of our proposed RP-DFT and other well-known techniques (equi-DFT, vari-DFT, and SWAT) based on Percentage Reduction, Recent-pattern biased error rate (Err_{RBP}), and RER from the real data.

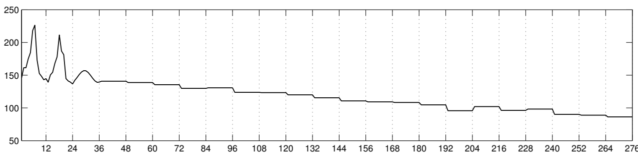

Figure 7: A reconstructed time series of the water usage data of 46 selected DFT coefficients from the original data of 276 data points, which is 83% reduction.

criterion is designed to measure the effectiveness of the algorithm after dimension reduction as following.

Definition 7 If X and $\tilde{X}$ are the original and reconstructed time series, respectively, *the recent-pattern biased error rate* is defined as

$$Err_{RPB}(X, \tilde{X}) = \mathbf{B} \cdot d_H^2(\hat{X}, \hat{\tilde{X}})$$
$$= \frac{1}{2} \sum_{i=0}^{\lfloor N/p \rfloor - 1} b(i) \left(\sqrt{\hat{x}_i} - \sqrt{\hat{\tilde{x}}_i} \right)^2, \qquad (10)$$

where **B** is a recent-pattern biased vector (which is a sigmoid function in our case).

Definition 8 If X and $\tilde{X}$ are the original and reconstructed time series, respectively and $Err_{RPB}(X, \tilde{X})$ is the recent-pattern biased error rate, then *the Reduction-to-Error Ratio* (RER) is defined as

$$RER = \frac{Percentage\ Reduction}{Err_{RPB}(X, \tilde{X})}. \qquad (11)$$

We compare the performance of our recent-pattern biased dimension reduction algorithm (RP-DFT) to equi-DFT, vari-DFT (with $k = 8$), and SWAT as we apply these algorithms on the mobile phone and water usage data. Table 1 shows the values of percentage reduction, recent-pattern biased error rate, and RER for each algorithm. It shows that SWAT has the highest reduction rates as well as the highest error rates in both data. For the mobile phone data, the values of the percentage reduction are the same for our RP-DFT and equi-DFT because R is exactly a half of the time series hence the sigmoid function is placed at the half point of the time series ($N/2$) that makes it similar to equi-DFT (in which the number of coefficients is exponentially decreased). The error rate of our RP-DFT is however better than equi-DFT by keeping more coefficients particularly for the "recent-pattern data" and fewer for older data instead of keeping more coefficients for just recent data and fewer for older data. As a result, RP-DFT performs with the best RER among others. For the water usage data, even though RP-DFT has higher error rate than equi-DFT, R is a short portion of the entire series thus RP-DFT is able to achieve much higher reduction rate, which results in a better RER.

In addition to the result of the performance comparison on the real data, we generate 100 synthetic data to further evaluate our algorithm compared to others. After applying each algorithm to these 100 different synthetic time series, Table 2 shows the average values of percentage reduction, recent-pattern biased error rate, and RER for each algorithm, where our proposed algorithm has the highest RER among others.

Algorithm	Percentage Reduction	Err_{RBP}	RER
RP-DFT	0.758	0.0209	**36.268**
equi-DFT	0.481	0.0192	25.052
vari-DFT	0.748	0.0385	19.429
SWAT	0.975	0.109	8.945

Table 2: Performance comparison of our proposed RP-DFT and other well-known techniques (equi-DFT, vari-DFT, and SWAT) based on the average Percentage Reduction, Recent-pattern biased error rate (Err_{RBP}), and RER from 100 synthetic data.

Conclusions and Future Work

In this paper, we have developed a novel dimension reduction framework for a human behavioral time series based on the recent pattern. With our framework, more details are kept for recent-pattern data, while older data are kept at coarser level. Unlike other recently proposed dimension reduction techniques for recent-biased time series analysis, our framework emphasizes on keeping the data that carries the most recent pattern (behavior), which is the most important data portion in the time series with a high resolution while retaining older data with a lower resolution. Our experiments on synthetic data as well as real data demonstrate that our proposed framework is very efficient and it outperforms other well-known recent-biased dimension reduction techniques. As our future directions, we will continue to examine various aspects of our framework to improve its performance as well as expand it to work with other types of time series.

Acknowledgments

We would like to express our gratitude and appreciation to Professor Xiaohui Yuan at UNT for his many constructive suggestions. This work is supported by the National Science Foundation under grants CNS-0627754, CNS-0619871and CNS-0551694.

References

Aggarwal, C.; Han, J.; Wang, J.; and Yu, P. 2003. A Framework for Clustering Evolving Data Streams. *In Proceedings of the 29th Very Large Data Bases Conference*.

Berberidis, C.; Aref, W. G.; Atallah, M.; Vlahavas, I; and Elmagarmid, A. 2002. Multiple and Partial Periodicity Mining in Time Series Databases *In Proceedings of 15th European Conference on Artificial Intelligence*.

Bulut, A.; and Singh, A.K. 2003. SWAT: Hierarchical Stream Summarization in Large Networks. *In Proceedings of the 19th International Conference on Data Engineering*.

Chen, Y.; and Dong, G. Han, J.; Wah, B. W.; and Wang, J. 2002. Multi-Dimensional Regression Analysis of Time-Series Data Streams. *In Proceedings of the 2002 International Conference on Very Large Data Bases (VLDB'02)*.

Elfeky, M. G.; Aref, W. G.; and Elmagarmid, A. K. 2004. Using Convolution to Mine Obscure Periodic Patterns in One Pass. *In Proceedings of the 9th International Conference on Extending Data Base Technology*.

Elfeky, M. G.; Aref, W. G.; and Elmagarmid A. K. 2005. Periodicity Detection in Time Series Databases. *IEEE Transactions on Knowledge and Data Engineering* 17(7):875-887.

Fu, T.; Fu, T-c.; Chung, F.l.; Ng, V.; and Luk, R. 2001. Pattern Discovery from Stock Time Series Using Self-Organizing Maps. *Notes KDD2001 Workshop Temporal Data Mining*, pp. 27-37.

Giannella, C.; Han, J.; Pei, J.; Yan, X.; and Yu, P.S. 2003 Mining Frequent Patterns in Data Streams at Multiple Time Granularities. *Data Mining: Next Generation Challenges and Future Directions*, H. Kargupta, A. Joshi, K. Sivakumar, and Y. Yesha, eds., AAAI/ MIT Press, 2003.

Hipel, K. W.; and McLeod, A. I. 1995. *Time Series Modelling of Water Resources and Environmental Systems*, Elsevier Science B.V.

Indyk, P.; Koudas, N.;and Muthukrishnan, S. 2000. Identifying Representative Trends in Massive Time Series Data Sets Using Sketches. *In Proceedings of the 26th International Conference on Very Large Data Bases*.

Keogh, E.; Chakrabati, K.; Pazzani, M.; and Mehrota, S. 2000. Dimensionality Reduction for Fast Similarity Search in Large Time Series Databases. *Knowledge and Information Systems*, 3(3)263-286.

Ma, S.; and Hellerstein, J. 2001. Mining Partially Periodic Event Patterns with Unknown Periods. *In Proceedings on the 17th International Conference on Data Engineering*.

Palpanas, T.; Vlachos, M.; Keogh, E.; Gunopulos, D.; and Truppel, W. 2004. Online Amnesic Approximation of Streaming Time Series. *In Proceedings of the 20th International Conference on Data Engineering*.

Perng, C.-S.; Wang, H.; Zhang, S.R.; and Parker, D.S. 2000. Landmarks: A New Model for Similarity-Based Pattern Querying in Time Series Databases. *In Proceedings on the 16th International Conference on Data Engineering*.

Phithakkitnukoon, S.; and Dantu, R. 2008. UNT Mobile Phone Communication Dataset. Available at http://nsl.unt.edu/santi/Dataset%20Collection/Data%20description/data_desc.pdf

Remagnino, P.; Foresti, G. L. 2005. Ambient Intelligence: A New Multidisciplinary Paradigm. *IEEE Transactions on Systems, Man, and CyberneticsPart A: Systems and Humans*, 35(1)1-6.

Yang, G. L.; and Le Cam, L. M. 2000. *Asymptotics in Statistics: Some Basic Concepts*, Berlin, Springer, 2000.

Yang, J.; Wang, W.; and Yu, P. 2000. Mining Asynchronous Periodic Patterns in Time Series Data. *In Proceedings on the 6th International Conference on Knowledge Discovery and Data Mining*.

Zhao, Y.; and Zhang, S. 2006. Generalized Dimension-Reduction Framework for Recent-Biased Time Series Analysis. *IEEE Transactions on Knowledge and Data Engineering*, 18(2)1048-1053.