World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:8, No:2, 2014

REDUCER – An Architectural Design Pattern for Reducing Large and Noisy Data Sets

Apkar Salatian

***Abstract*—**To relieve the burden of reasoning on a point to point basis, in many domains there is a need to reduce large and noisy data sets into trends for qualitative reasoning. In this paper we propose and describe a new architectural design pattern called REDUCER for reducing large and noisy data sets that can be tailored for particular situations. REDUCER consists of 2 consecutive processes: *Filter* which takes the original data and removes outliers, inconsistencies or noise; and *Compression* which takes the filtered data and derives trends in the data. In this seminal article we also show how REDUCER has successfully been applied to 3 different case studies.

***Keywords*—**Design Pattern, filtering, compression.

I. Introduction

IN many domains there is a need to reduce large and noisy data sets. Reduction of such data may typically involve pre-processing of the data to remove outliers, inconsistencies or noise. This filtered data would then be compressed into trends so that interpretation becomes easier in the form of qualitative reasoning. Such a common approach lends itself to the development of an architectural design pattern.

Architectural design patterns are considered templates or descriptions of how to solve a problem that can be used in many different situations. They are general reusable solution to commonly occurring problems.

In this seminal paper we propose and describe the REDUCER architectural design pattern for reducing large and noisy data sets. REDUCER consists of 2 consecutive processes: *Filter* which takes the original data and removes outliers, inconsistencies or noise; and *Compression* which generates trends in the underlying data to make interpretation easier in the form of qualitative reasoning.

The structure of this paper is as follows. Section II proposes and describes the REDUCER architectural design pattern to reduce large and noisy data sets into trends. Section III describes how the REDUCER architectural design pattern has been applied to 3 different case studies. A discussion of our work is given in Section IV and final conclusions are given in Section V.

II. Reducer Design Pattern

Fig. 1 depicts the architecture of our REDUCER design pattern. Data is initially filtered to remove outliers, inconsistencies and noise. The filtered data is then compressed into trends by a second process and the results allows qualitative reasoning for interpretation.

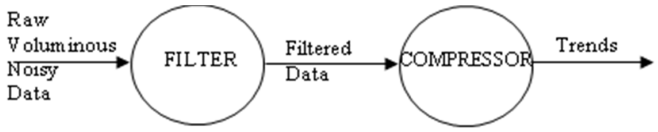

Fig. 1 The REDUCER Design Pattern

III. Application of the Reducer Design Pattern

We will demonstrate the application of the REDUCER architectural design pattern to 3 different case studies: deriving trends in a heart rate trace using the random selection approach; deriving trends in building data using an agglomerative approach; and deriving trends in a blood pressure trace using wavelet analysis.

A. Random Selection Approach

Reference [1] uses the random selection approach to derive trends in a heart rate trace. In the random selection approach, to reduce a data set D containing n rows of information to one containing k rows of information, a set S_k is formed consisting of k numbers selected at random from the set S given by:

$$S = \{x \in \mathrm{N} \mid 1 \le x \le n\}$$

Then, our reduced set, D_R, will be given by:

$$D_R = D(S_k, :)$$

That is, D_R is a data set having the same number of columns as D, and the i^{th} row of D_R will be the j^{th} row of D if j is the i^{th} element of S_k.

To demonstrate the random selection approach, consider a data set, *A1*, which is the heart rate of a patient recorded at a frequency of one value every minute. The data set has 3785 rows, one row for each minute in this time interval. The graph representing this data is shown in Fig. 2.

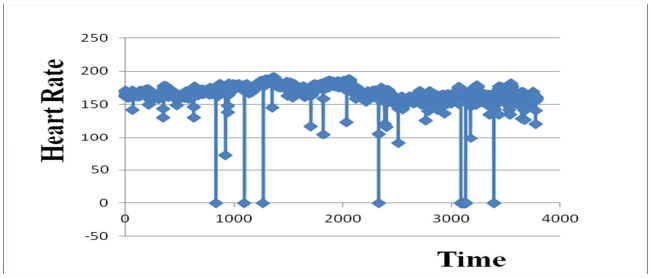

Fig. 2 Original Heart Rate Trace

A.Salatian is with the School of Information Technology and Computing at the American University of Nigeria, Yola Bypass, PMB 2250, Yola, Nigeria (phone +234 08052000507; e-mail: apkar.salatian@aun.edu.ng).

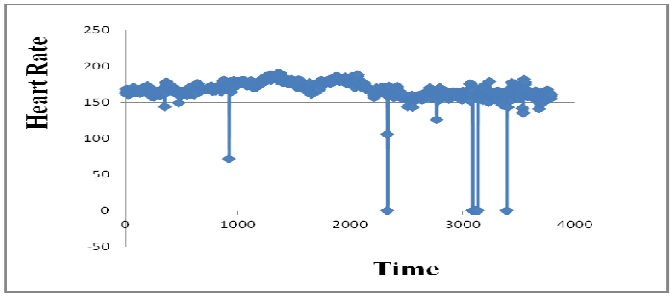

Fig. 3 Data Set reduced to 2000 rows

To reduce the data set to, say, 2000 rows, we randomly select 2000 rows from the 3785 rows in *A1* to give a data set *A2*. Here our algorithm randomly generates 2000 random numbers between 0 and 1, multiplies each number by 3785 and stores the result in an array *A1*. This process can be considered the *filter* process of REDUCER because it will suppress outliers. The array is then sorted by index to form a sorted array *A2* with 2000 values ranging from 0 to 3785. Then, a new matrix *A3* is generated containing, for each entry, *x*, in *A2*, the x^{th} row of *A1*. In other words, if *x* is the i^{th} entry of *A2*, then the x^{th} row of the original data set will become the i^{th} row of *A*. The reduced data set, *A3*, is shown in Fig. 3.This process can be considered the *compressor* process of REDUCER because the data set has been compressed from 3785 points to 2000 points.

B. Agglomerative Approach

Building operators are confronted with large volumes of continuous noisy data from multiple environmental sensors which require interpretation. The ABSTRACTOR [2] system summarizes historical environmental sensor data for reporting and building performance assessment using an agglomerative (merging) approach. We shall describe how ABSTRACTOR implemented each of the processes of the REDUCER design pattern

Initially the data needs to be filtered to get rid of non-significant events in environmental monitoring data. ABSTRACTOR uses an average filter which involves a moving window which is centered on a point x_n and if the window is of size *2k+1* the window contains the points x_{n-k} to x_{n+k}. This process can be considered the *filter* process of REDUCER because by always choosing the average value in the window as the filtered value it removes all the very short duration spikes from the outdoor temperature data whilst revealing the short duration trends hidden in the raw data.

The algorithm derives trends in the filtered by following two consecutive sub-processes called *temporal interpolation* and *temporal inferencing*. *Temporal interpolation* takes the filtered data *a*nd generates simple intervals between each consecutive data point. *Temporal inferencing* takes these simple intervals and tries to merge them into longer trends – this is achieved by using 4 variables: *diff* which is the variance allowed to derive steady trends, *g1* and *g2* which are gradient values used to derive increasing and decreasing trends and *dur* which is used to merge intervals based on the duration of the middle interval when 3 intervals are being merged. Temporal Inferencing rules to merge 2 meeting intervals (Δ_{H2}) and 3 meeting intervals (Δ_{H3}) use the 4 variables to try to merge intervals into larger intervals until no more merging can take place. This process can be considered the *compressor* process of REDUCER because the filtered data points have been compressed into underlying trends in the data.

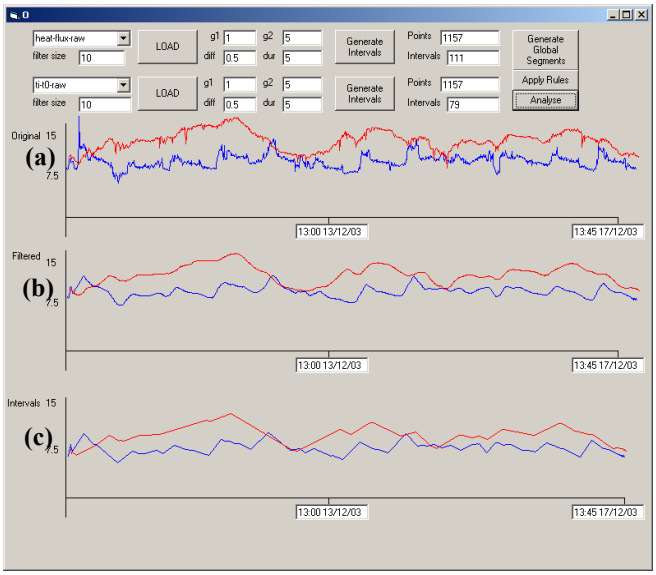

Fig. 4 ABSTRACTOR applied to environmental data

Fig. 4 shows the results of ABSTRACTOR. The data was 12179 minutes worth of continuous data (see Fig. 4 (a)). The data was the heat-flux into a wall and the difference in internal and external temperature (ti-t0) measurements; the sampling frequency of the signals is one data item every 15 minutes. The application of the average filter (k=10 filter provides a running five and a quarter hour running average) is shown in the middle graph (b) and the final intervals (trends) generated are shown in the bottom graph (c).

C. Wavelet Analysis

In [3] the authors use wavelet analysis to derive trends in a blood pressure trace. A wavelet is a time-series mathematical function used to divide a given function or continuous-time signal into different scale components. For a discussion of different wavelets functions the reader is advised to read [4].

The detection and estimation of trends in the presence of stochastic noise arises in many data sets. Wavelet analysis is a transformation of a time series of data in which we obtain two types of coefficients: wavelet coefficients and scaling coefficients - these are sometimes referred to as the *mother* and *father wavelet coefficients* respectively. The wavelets are scaled and translated copies (*father wavelets*) of a finite-length or fast-decaying oscillating waveform (*mother wavelet*).

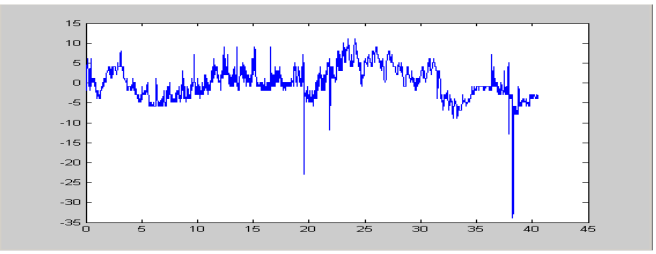

Fig. 5 Original Blood Pressure Trace

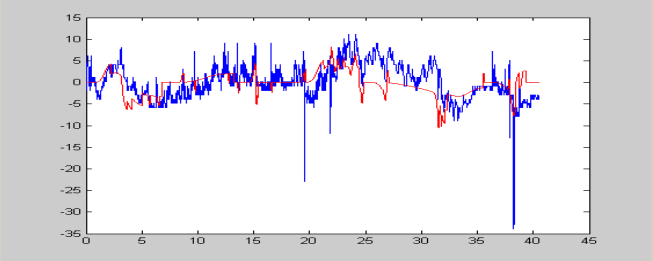

Fig. 6 Trend Composition

In [3] the authors tried to derive trends in a blood pressure trace taken from an Intensive Care Unit in the United Kingdom. Fig. 5 shows the waveform of the blood pressure signal – the frequency of the data is 15 - 20Hz. It can be seen that the data is noisy due to clinically insignificant events such as line flushes and taking of blood samples from the patient. This data had a mother wavelet estimation process applied to it to generate the father wavelet coefficients. This is considered the filter process of REDUCER because it dilates the signal and suppresses any noise in the signal. The resultant data is then put through a Continuous Wavelet Transform (CWT) process to find the amplitude of the frequency components in the data at different times – this derives the underlying trends in the data. The trends are shown in Fig. 6 and are represented in red. The CWT process can be considered the *compressor* process of REDUCER.

IV. Discussion

In all our case studies we performed negative filtering to remove or suppress outliers, inconsistencies and noise. We have shown that filtering can take different forms such as using random numbers, averaging and dilation. Other filtering techniques include using a median filter, low-pass filter, and high-pass filter [5].

Likewise, we have shown that compressing the data into trends can also take on different forms. Compression of data relieves the burden of reasoning on a point to point basis by reducing large and noisy data sets into trends for another higher level process to perform qualitative reasoning.

V. Conclusions

The compression of large and noisy data sets is non-trivial – one approach is to have an architectural design pattern which can be tailored to a particular situation. We believe that the REDUCER architectural design pattern is a step in the right direction in the compression of large and noisy data.

We have shown that the REDUCER architectural design pattern can be tailored and applied to different domains which have the same issues associated with the compression of data.

Since this is a seminal paper, there is no direct comparison with other work. Consequently we hope that we have provided researchers dealing with large and noisy data sets an architectural design pattern that is generic enough to be used within any domain and discipline. Our future work will be to develop a tool for this architectural design pattern which will lend itself for reuse.

References

[1] Nsang, A., Salatian, A. "Data Reduction of ICU Data using a Random Selection Approach", International Journal of Advanced Science and Technology, Vol. 55, 2013, pp. 81-88,.

[2] Salatian, A., & Taylor, B., "ABSTRACTOR: An Agglomerative Approach to Interpreting Building Monitoring Data", Journal of Information Technology in Construction, Vol. 13, May 2008, pages 193-211.

[3] Salatian, A. & Adepoju, F. "In Praise of Wavelets – 3 Disparate Case Studies", The 3rd International Multi-Conference on Complexity, Informatics and Cybernetics: IMCIC 2012, Volume 1, pages 36 – 40. Orlando, Florida, USA, March 25th - 28th, 2012.

[4] Singhal A, Singh RP, Tenguria M, "Comparison of Different Wavelets for Watermarking of Colored Images", 2011 IEEE 3rd International Conference on Electronics Computer Technology (ICECT 2011), Kanyakumari, India, pp. 187 – 191.

[5] Salatian, A. & Oborkhale, L., "Filtering of ICU Monitor Data to Reduce False Alarms and Enhance Clinical Decision Support", International Journal of Bio-Science and Bio-Technology, Volume 3, Number 3, June 2011, pages 49-55.