

Validating the FTI Design Pattern for the Interpretation of Voluminous High Frequency and Noisy Data Sets

Apkar Salatian

School of Information Technology and Communications, American University of Nigeria, Yola Bypass,
PMB 2250, Yola, Nigeria
apkar.salatian@aun.edu.ng

Abstract. In this article we propose and validate the FTI design pattern for the interpretation of voluminous high frequency and noisy data sets. The FTI design pattern consists of 3 consecutive processes: Filter which takes the original data and removes outliers and noise; Trender which abstracts the filtered data into trends; and Interpretation which uses tailored knowledge bases to perform qualitative reasoning of the trends to provide an analysis of the original data. An empirical case study is presented to demonstrate the FTI design pattern and then by performing pattern mining on six further systems in the way we have suggested herein, we argue that our FTI design pattern has been validated.

Keywords: Design Pattern, filtering, data compression, interpretation

1. Introduction

In many domains there is a need to interpret high volumes of high frequency noisy data. Interpretation of such data may typically involve pre-processing of the data to remove outliers or noise. Due to the volume of data it would be computationally expensive to reason quantitatively on a point to point basis so it would be better to reduce the volume of data by abstracting the data into trends so another process can perform qualitative reasoning to provide analysis of the original data. Such a common approach lends itself to the development of a design pattern.

Design pattern describes a proven design solution to a recurring programming problem, and they have become very widely used in software development. They place particular emphasis on the context and forces surrounding the problem, and the consequences and impact of the solution. They constitute a base of experience for building reusable software. A design pattern is rendered as a parameterized collaboration. As a collaboration, a pattern provides a set of abstractions whose structure and behaviour work together to carry out some useful function. The collaboration's parameters name the elements that a user of this pattern must bind. Therefore, design patterns are a template that you use in a particular context by supplying elements that match the template parameters [1].

In this paper we describe and validate the *FTI (Filter – Trender – Interpretation)* design pattern for interpreting voluminous high frequency and noisy data sets. FTI consists of 3 consecutive processes: *Filter* which takes the original data and removes outliers and noise; *Trender* which takes the filtered data and derives trends (abstractions) in the data – this acts as a form of data compression; and *Interpretation* which uses tailored knowledge bases to perform qualitative reasoning of the trends to provide an analysis of the original data.

2. The FTI Design Pattern

Figure 1 depicts the FTI design pattern. The noisy and voluminous data is initially filtered to get rid of outliers and noise. Given continuous data (in some cases up to one value every second), it is computationally expensive to reason quantitatively with each data value on a point to point basis. *Trender* is the abstraction of

the filtered data generated by the filter process into trends in which data as *steady*, *increasing* and *decreasing*. One may also be interested in the rate of change e.g *rapidly increasing*, *slowly decreasing* etc. One must determine from the data where the abstractions begin and end. Using tailored knowledge bases, these trends are finally interpreted by a third process which provides an analysis of the original data.

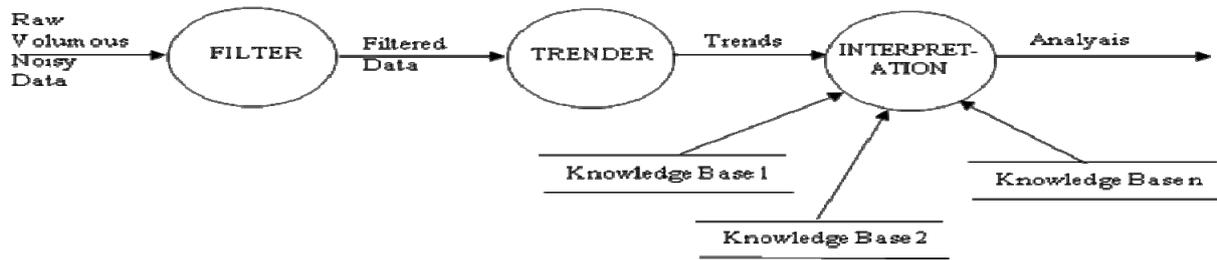


Figure 1 – The FTI Design Pattern

3. Application of the FTI Design Pattern

We will demonstrate the application of the FTI design pattern for the interpretation of Intensive Care Unit (ICU) monitor data such as the heart rate and blood pressure which can have a frequency of up to 1 value every second – this creates information overload for medical staff who need to interpret the data to evaluate the state of the patient.

3.1 Filter

Due to nature and frequency of the data, a median filter was chosen to remove the noise like clinically insignificant events from the physiological parameters. The median filter 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} . By always choosing the median value in the window as the filtered value, it will remove transient features lasting shorter than k without distortion of the base line signal; features lasting more than that will remain. Clinically insignificant events which cannot be removed at this stage will be dealt with by the Interpretation process.

3.2 Trender

The algorithm for identifying trends involves following two consecutive sub-processes called temporal interpolation and temporal inferencing. Temporal interpolation takes the filtered data and generates simple intervals between consecutive data point. Temporal inferencing is an agglomerative algorithm that takes these simple intervals and tries to generate larger trends by merging similar intervals – this is achieved 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 3 intervals based on the duration of the middle interval. 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. For further discussion of the algorithm the reader is advised to read [3].

3.3 Interpretation

Interpretation involves interpreting overlapping trends. This process serves two purposes: removal of clinically insignificant events; and the generation of graphical summaries and patient state assessments in the form of clinical conditions the patient encountered and the outcome of therapies.

All events (clinically insignificant, clinical conditions and outcomes of therapies) that typically occur in the ICU are represented as user defined trend templates (ontology) in their respective knowledge bases. The schema that allows trend templates to be made up encompasses three types of knowledge: temporal, differential and taxonomical. Temporal knowledge allows temporal reasoning; interval-based and point-based reasoning. Since several clinical conditions may be described by the same patterns, differential knowledge is be used to eliminate possibilities and hence prevent unnecessary reasoning – this can include the patient's history. Within the trend templates there is also taxonomical knowledge - since several clinical conditions have similar attributes, this enables us to represent them as a hierarchy of classes and subclasses. A schema

has been developed which allows inheritance of attributes. Such a representation allows more abstract clinical conditions to be identified - if a specific instance of a clinical conditions cannot be identified then the more general class of clinical condition to which it belongs is more likely to describe the data.

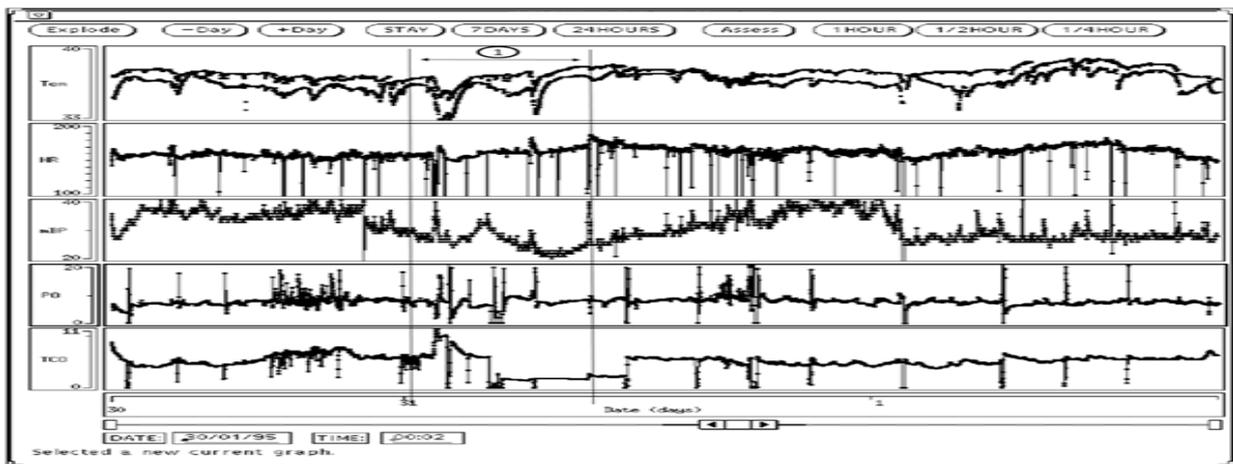


Figure 2 – Original Signals

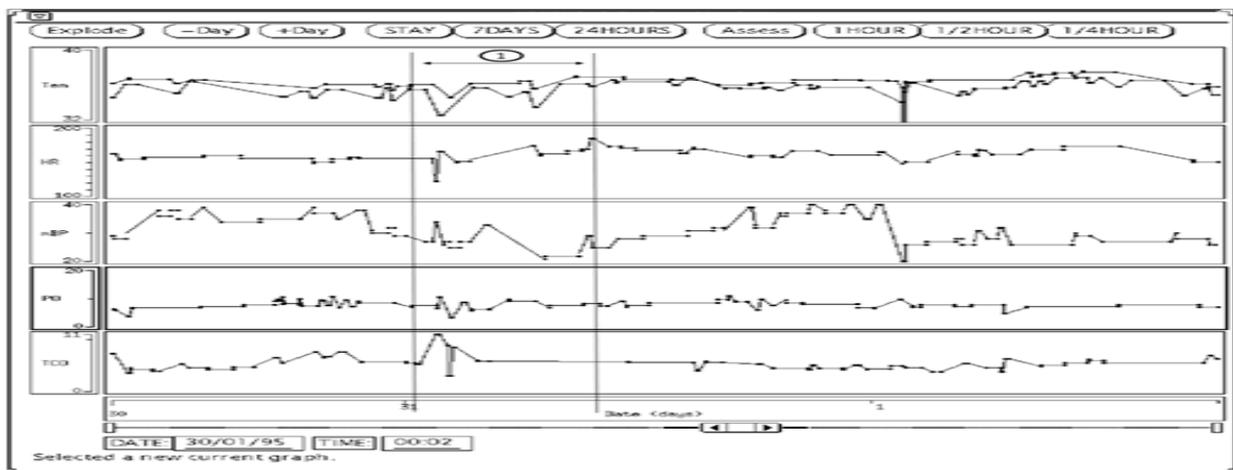


Figure 3 – Trends derived from the Trender process

Pattern-matching algorithms which exploit the trend templates in the knowledge bases have been implemented to deal with the typical events that are encountered in the ICU. Belief revision algorithms have been implemented to identify clinically significant events (clinical conditions and outcome of therapies) which are analogous to the way clinicians interpret clinical data. An associational-reasoning algorithm has been implemented to identify clinically insignificant events. These algorithms can deal with different signals simultaneously. Pattern-matching is made by comparing parameters quantitatively against numerical values and symbolic ranges, and qualitatively against the direction of change and expectations. For further discussion of the algorithm the reader is advised to read [2].

As an example, consider a three day data set taken from a neonatal ICU in the UK from 00:01 on 30 January 1995 to 23:59 on 1 February 1995. It is known that this patient experienced a hypovolaemia 'between 12am and 4am on 31 January 1995'. Figure 2 depicts the original signal and figure 3 depicts a graphical summary of the temporal intervals generated for each parameter. Note that in the graphs, *Tem* represents the (higher) core and the (lower) peripheral temperature. *HR* represents the Heart Rate, *mBP* represents the mean Blood Pressure, *PO* represents the Partial Pressure of Oxygen and *TCO* represents the Partial Pressure of Carbon Dioxide.

All but two of the clinically insignificant events found were correct.

For the clinical condition interpretation, ASSOCIATE identified the hypovolaemia at various intervals within the time points specified but not for its full duration. The immediate class that hypovolaemia inherits, infant shock, has been identified for lasting for greater durations within the time points specified – this is a

successful application of taxonomical knowledge in the knowledge bases. Note, the period 12am to 4am on 31 January 1995 is marked with by the interval labelled 1 on the original data set (figure 2) and by the interval labelled 1 on the graphical summary (figure 3). The reason hypovolaemia was not fully captured was that the heart rate did not increase as one would expect but remained steady. The expert agreed that, though not always specific in its identification, ASSOCIATE recognised all clinical conditions in the data set i.e no clinical conditions were missed.

For the therapy interpretation, all the results are incorrect because the changes were not significant or the assertion was entered at an approximate time after the therapy was made.

4. Discussion

By performing pattern mining in the form of tailoring and adapting different systems to perform filtering, interpreting and reporting we will now further validate the FTI design pattern.

The ABTRACTOR [4] system uses the FTI design pattern to summarise historical building monitor data from multiple environmental sensors for interpretation and building performance assessment. Initially the data is filtered by an average filter to get rid of non-significant events – here all the very short duration spikes from the data were removed whilst revealing the short duration trends hidden in the raw data. The *Trender* process for ABTRACTOR is exactly the same as the agglomerative approach used for our case study - for a discussion of this algorithm applied to building monitor data the reader is advised to read [5]. Interpretation involves splitting the overlapping trends into *global segments* in the spirit of [6] – here a change in the direction of change of one (or more) channels or a change in the rate of change of one (or more) channels contributes to a split in the trends creating a global segment. The algorithm for interpretation then involves applying rules from the knowledge bases to the global segments to identify particular events. If rules are true over adjacent global segments then one can determine when the environmental event started and ended.

BT-45[7] generates natural language textual summaries of continuous physiological signals and discrete events from a Neonatal Intensive Care Unit. BT-45 could be adapted to use the FTI design pattern. The first stage of BT-45 is *Signal Analysis* which extracts the main features of the physiological time series – this fulfils the role of the Filter process of FTI. The next stage of BT-45 is *Data Interpretation* which performs some temporal and logical reasoning to infer more abstract medical observations and relations from the signal features – this could be considered as the *Trender* process of FTI since it is a form of data compression. The next 2 stages are *Document Planning* which selects the most important events from earlier stages and groups them into a tree of linked events and finally, *Microplanning and Realisation* which translates this tree into coherent text for reporting - collectively they could be considered the Interpretation process of FTI since an analysis of the abstractions is being made.

SUMTIME-MOUSAM [8] is a text generator that produces textual marine weather forecasts for offshore oilrig applications. It uses a subset of the processes of BT-45 and also follows the FTI design pattern. The architecture of SUMTIME-MOUSAM follows 3 processes: *Document planning*, *Micro planning* and *Realization*. Document planning is responsible for selecting the ‘important’ data points from the input data and to organize them into a paragraph – this is a form of filtering. Micro planning is responsible for lexical selection and ellipsis – like the *Trender* process, this is a form of abstraction of the filtered data. Realization is essentially responsible for ordering of the phrases in the output and also to perform punctuation tasks – this is analogous to the Interpretation process of FTI.

A similar approach is taken by [9] to generate textual summaries of geo-referenced data based on spatial reference frames. From the initial data basic events are generated (filtered out) by a data analysis process which is then abstracted into higher level concepts. The final stage is to report these messages in sentence form for textual summarization. These processes are also analogous to FTI.

[10] extract extra knowledge from clickthrough data of a Web search engine to improve web-page summarization. Among the 3,074,678 Web pages crawled, the authors removed those which belong to “World” and “Regional” categories, as many of them are not in English – this filtering resulted in 1,125,207 Web pages, 260,763 of which are clicked by Web users using 1,586,472 different queries. Three human evaluators were employed to summarize (abstract) these pages. Each evaluator was requested to extract the

sentences which he/she deemed to be the most important ones for a Web page. A graph (interpretation) of the precision of the query terms was finally reported.

[11] presented a case-based reasoning approach to activity recognition in a smart home setting. An analysis was performed on scalability with respect to case storage, and an ontology-based approach was proposed for case base maintenance – this could lend itself to the FTI pattern. Firstly to create a cut-down (filtered) case base a reduction was made by firstly using a simple statistical technique, and then by semantically linking the case solutions with corresponding case features – this could be considered a form of abstraction. The case solutions were analysed and some had their accuracy reduced while others had theirs increased – this is considered a form of interpretation.

5. Summary and Conclusions

The interpretation of voluminous high frequency and noisy data is non-trivial – one approach is to have a design pattern which can be tailored and applied to different domains which have the same issues associated with the interpretation of the data.

We have shown that research into trying to interpret high volume noisy datasets do not actually follow any particular design pattern or framework - they just tell us about the ‘tactics’ they have employed in order to process such data. By conducting a detailed empirical study of the author’s former research endeavours and pattern mining six further systems, we believe that we have successfully argued that our FTI design pattern allows systems to be adapted at a much higher level of abstraction to facilitate the interpretation of voluminous high frequency and noisy data leaving the tactics to the individual processes.

6. References

- [1] Booch, G., Rumbaugh, J., Jacobson, I., [Unified Modeling Language User Guide, The (2nd Edition)], page 32, Addison-Wesley Object Technology Series (2005).
- [2] Salatian, A., “Interpreting Historical ICU Data Using Associational and Temporal Reasoning”, 15th IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2003), Sacramento, USA, 442-450 (2003).
- [3] Salatian, A. Hunter, J. R. W., “Deriving trends in historical and real-time continuously sampled medical data”, *Journal of Intelligent Information Systems*, 13, 47-74 (1999).
- [4] Salatian, A., Taylor, B., “ABSTRACTOR: An Agglomerative Approach to Interpreting Building Monitoring Data”, *Journal of Information Technology in Construction*, 13, 193-211 (2008).
- [5] Salatian, A., & Taylor, B., “An Agglomerative Approach to Creating Models of Building Monitoring Data”, *Proceedings of the eighth IASTED International Conference on Artificial Intelligence and Soft Computing*, Marbella, Spain, 167-172 (2004).
- [6] DeCoste, D., “Dynamic Across-Time Measurement Interpretation”, *Artificial Intelligence* 51, 273-341 (1991).
- [7] Portet, F, Reiter, E., Gatt, A., Hunter, J.R.W., Sripada, S., Freer, Y., Sykes, C., "Automatic Generation of Textual Summaries from Neonatal Intensive Care Data", *Artificial Intelligence*, 173, 789-816 (2009).
- [8] Sripada, S., Reiter, E. Davy, I., “SumTime-Mousam: Configurable marine weather forecast generator”, *Expert Update*. 6(3), 4-10 (2003)
- [9] Turner, R., Sripada, S., Reiter, E. and Davy, I., “Using spatial reference frames to generate grounded textual summaries of georeferenced data”, *Proceedings of the Fifth International Natural Language Generation Conference*, Salt Fork, Ohio, 16-24 (2008).
- [10] Sun, J-T., Shen, D., Zeng, H-J., Yang, Q., Lu, Y., Chen, Z., “Web-page summarization using clickthrough data”, *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, Brazil, 194-201 (2005).
- [11] Knox, S., Coyle, L., Dobson, S., “Using ontologies in case-based activity recognition”, in *23rd Florida Artificial Intelligence Research Society Conference*, AAAI Press, 336-341. (2010).