Smart decimation method for fusion research data

Rodrigo Castro[a], Jesús Vega[a]

[a]*CIEMAT Fusion program. Avda. Complutense 40. Madrid. Spain*

***Abstract*— New fusion research experiments will generate massive experimental data. In this context, fusion research appears in the scope of the big data where both search and access functions require new approaches and optimizations.**

One common data access functionality is decimation. It allows to retrieve a limited number of points of the total available. One classical mechanism of decimation is downsampling by an integer factor called step. It works by selecting a value every 'step' number of values. The main characteristic of classical decimation is that the selected values are uniformly distributed along the total. However, in case of time evolution experimental signals, the relevancy of data is not uniform. There are some intervals where the provided information is more complex and richer, and usually more interesting from the user point of view.

This contribution presents a new data decimation technology for unidimensional time evolution signals where the limited number of accessed points are distributed following the criterion of data interest level. The new method implements, on one hand, a heuristic function which is able to determine the level of interest of an interval based on its data characteristics, and on the other hand, a selection algorithm where points are distributed based on weighted intervals.

Keywords—decimation;acess;access;fusion;search;archiving

I. Introduction

New fusion energy research experiments will generate massive experimental data. For example, ITER (International Thermonuclear Experimental Reactor) will have above one million of variables coming from control signals and diagnostic systems. Some of these variables will produce data during long pulse (about 30 minutes) experiments, while other will generate data continuously. Just to have a clearer idea of the scenario, ITER estimates more than 30 GBytes/second of data flow during experiment pulses. In this context, traditional complete access to archived data is very expensive in computing resources and time, and new data access that will help researchers to locate and find useful data are required.

In this work a new decimation method developed in CIEMAT (Spanish Energy Research Centre) for time evolution signal data is presented. The main objective of the new method is to obtain quickly low resolution and high similarity views of archived data. To achieve it, it is necessary to improve the classic 1-of-n decimation method that for high decimation factors produces very poor views of the original signal.

II. The Smart Decimation Method

The new decimation method, as other decimation methods, receives as input both a variable that represents a time evolution signal data (S) to decimate, and the number of points (N) which are expected. The method produces as output a sequence of N points (time and value) that have been selected from original signal data (S). The main characteristic of the presented method is that the output points are not homogeneously distributed (as other decimation methods); they are distributed proportionally to the level of interest of the signal along the time. This means that output points are more concentrated in more interesting time areas of the signal.

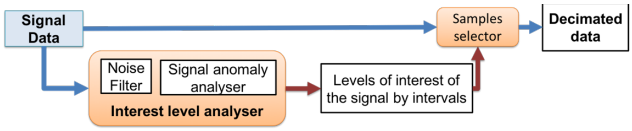


Fig 1 Architecture diagram of the smart decimation method. It includes two phases: Interest level analyzer and samples selector

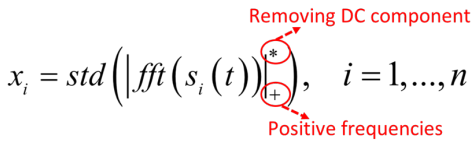
As it is shown in Fig 1, the method is implemented with a two-step algorithm. The first step measures the level of interest along the signal and it is implemented into the "Interest level analyser" component. The second step selects N values from the original signal based on the level of interest metadata and is implemented into the "Samples selector" component. These two steps are independent, so the interest level metadata can be created and archived while the original signal data are archived, and the selection of points will be performed for every data access with smart decimation.

A. Interest level analyser

The interest level analyser component is responsible of measuring the level of interest assigned to a portion of a time evolution signal. One valid approach to this objective is to use an algorithm that can measure the level of change in the behaviour of a time evolution signal. In our implementation we have used the core of an anomaly detention algorithm that has been successfully applied in the JET disruption detector named APODIS [1,2] which is based on the "Standard Deviation of Fourier Spectrum" method [3].

The standard deviation of Fourier spectrum method, Fig 2, processes a signal by regular intervals and calculates the anomaly level for every interval. The method has the following steps:

1. Calculating the Fourier Spectrum components of the interval
2. Removing the continuous component
3. Removing the negative components
4. Calculating the standard deviation of the remaining components



$$x_i = std\left(\left|fft\left(s_i(t)\right)\right|_{(*)}^{(+)}\right), \quad i = 1,\ldots,n$$

Fig 2 Formula of the standard deviation of Fourier spectrum

The output of this processing is a new signal metadata that represents the level of interest of the signal data along regular intervals.

One of the main motivations to select this method is its successful as anomaly detection algorithm applied to fusion experiment data. An additional motivation is its compatibility with real time implementation. This method can analyse and produce metadata in real time while the processed signals are acquired, and it is not necessary to wait until the signal is complete.

As it is shown in Fig 1, previous to the anomaly measurer, a noise filter has been included. It implements an exponential moving average algorithm [4], explained in Fig 3.

$$S(t) = \begin{cases} Y(0) & t = 0 \\ \alpha Y(t) + (1-\alpha) S(t-1) & t > 0 \end{cases}$$

Fig 3 Exponential moving average function

The main characteristic of this low pass filter is that it has a past memory forget mechanism, so it preserves changes in the behaviour of the time evolution signal. It is also important to remark that this algorithm can be fully implemented in real time.

B. No uniform samples selector

The objective of this component is to distribute the selected samples proportionally to a set of weights that, in this case, corresponds to the interest level metadata that were measured and persisted previously at regular time intervals.

In the description of the sample selection algorithm the following elements are used:

- Request interval: Interval of the decimated data request.
- Selection interval: Interval that is being considered to make a decimated selection of samples in the current step of the algorithm.
- Metadata interval: Interval in the original signal that corresponds to a metadata (interest level) value.

The steps of the algorithm are:

1. Selection interval = First metadata interval
2. While (number of decimated samples < N) and (selection interval into request interval)
   a. Number of samples to consider in the selection interval = (sum of weights in selection interval / sum of weights in request interval) * number of decimated samples left.
   b. Selecting the number of samples from the original signal to consider into the selection interval
   c. Shifting the selection interval to the next metadata interval

III. Some Results

The new decimation method has been tested with TJ-II (flexible Heliac TJ-II stellerator located at CIEMAT) experiment data. To test it, a typical 1-of-n decimation function (take 1 and jump n) has been compared with the new decimation method. The comparation has been done over a set of commonly used signals from different diagnostics and with different characteristics of sampling rate and noise. The Table 1 shows the list of used signals.

Table 1 List of TJ-II signals that have been used in new comparation test

Signal name	Description	Sapling rate
BOL1	Bolometer signal	100 KHz
DENCM0	Electron density	1 MHz
ECE10	Electron Cyclotron Emission	100 KHz
RX105	Soft Ray-X	78 KHz

Because the sort length of TJ-II pulses (0,5 seconds), long signals have been built concatenating the data of a set of pulses. Specifically, experiment data from pulse 46000 until pulse 46100 have been considered.

In order to obtain a comparative between the 1-of-n and the new decimation method, similarity between the original signal and decimated signal has been evaluated. To measure the similarity concept between two signals, two parameters have been taken into account: Euclidean distance and correlation coefficient. The Euclidean distance between two signals has been defined as the sum of the Euclidean distance between synchronous samples (samples with the same timestamp). Both analysed parameters, Euclidean distance and correlation coefficient, require having same time interval (same start and end timestamp), same number samples and synchronous samples (samples with the same timestamp). This is why, for this test, the decimated signals have been resampled to the timestamps of the original signal samples.

Results of the comparative are presented in Table 2. They show very significant improvement in both coefficients for all tested signals.

Table 2 Results of decimation comparation between 1-of-n and smart decimation methods

Signal	N. Samples	D. Method	E. Distance	Corr. Coef.
BOL1	6619136	1-of-n	107.5474	0.6727
		Smart	**67.2390**	**0.8914**
DENCM0	13107200	1-of-n	161.0769	0.9969
		Smart	**109.7335**	**0.9986**
ECE10	6205440	1-of-n	364.3474	0.5882
		Smart	**227.3609**	**0.8837**
RX105	5049984	1-of-n	187.5517	0.7732
		Smart	**99.0643**	**0.9421**

Apart from the test result numbers, the differences between the two compared decimation methods are more evident from a visual point of view. The Fig 4 and Fig 5 present visual comparatives of the two decimation methods for 1000 decimated points in a portion of the signals BOL1 and DENCM0 respectively. In both figures, the first graph shows the original signal, the second graph shows the plot of the 1-of-n decimated signal, the third graph shows the values of the interest level analysis and fourth graph shows the plot of the smart decimated signal.

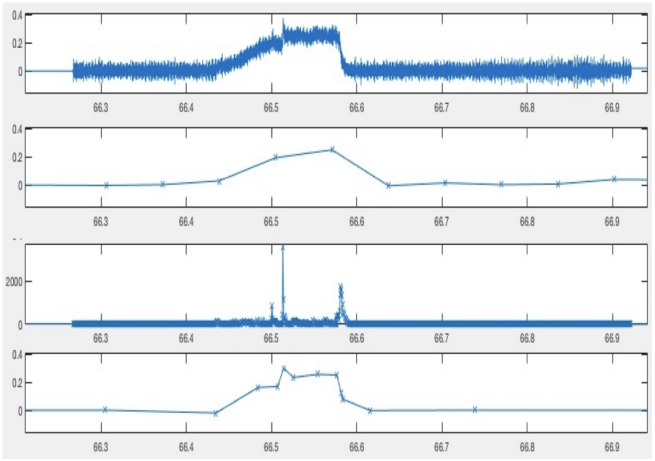


Fig 4 Visual comparative of 1-of-n and smart decimation method in a portion of BOL1 signal. The included graphs are (from top to down): original signal, 1-of-n decimated signal, levels of interest, smart decimated signal.

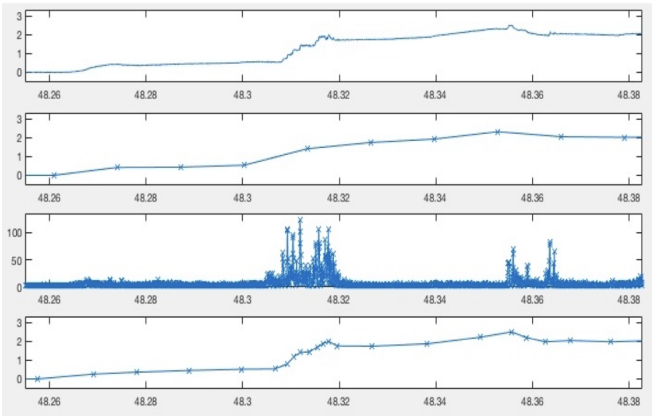


Fig 5 Visual comparative of 1-of-n and smart decimation method in a portion of DENCM0 signal. The included graphs are (from top to down): original signal, 1-of-n decimated signal, levels of interest, smart decimated signal

Both figures show some interesting results. Visually, the similarity between the original signal and the decimated signal is higher in case of smart decimation method (graphs 1 and 4) than 1-of-n method (graphs 1 and 2). Graphs 3 and 4 show how smart decimation works, concentrating decimation points in intervals with higher level of interest. Graphs 1 and 3 show how higher levels of interest in the original signal are located in intervals where the signal presents changes in its behaviour. In case of Fig 5, that corresponds to signal BOL1, it is interesting to remark the high level of noise in the original signal (graph 1).

IV. Conclusion

In order to improve fast and efficient data access to big data fusion experiments, a new smart decimation method for time evolution signals has been developed in CIEMAT. The method uses a signal anomaly analyser to calculate the level of interest by time intervals of the original signal and distributes decimation points based on it.

The smart decimation method has been successfully tested with TJ-II experimental signals. The test results show a very significant improvement as compared to the classic 1-of-n decimation method, in the similarity of output decimated signals and original signals. The new method can help researchers to quickly get useful low-resolution views of archived big data fusion experiments and to locate signals and zones of their interest.

V. References

1. J. M. López et al., "Implementation of the Disruption Predictor APODIS in JET's Real-Time Network Using the MARTe Framework," in IEEE Transactions on Nuclear Science, vol. 61, no. 2, pp. 741-744, April 2014
2. J.Vega, S. Dormido-Canto, J. M. López, AndreaMurari et al. , "Results of the JET real-time disruption predictor in the ITER-like wall campaigns", Fusion Engineering and Design, Volume 88, Issues 6–8, October 2013, Pages 1228-1231
3. G.A. Rattá,J. Vega. Murari, M.Johnson et al. ,"Feature extraction for improved disruption prediction analysis at JET", Review of Scientific Instruments. 79, 10F328 (2008).
4. J. E. Everett, "The Exponentially Weighted Moving Average Applied To The Control And Monitoring Of Varying Sample Sizes", WIT Transactions on Modelling and Simulation, Volume 51, Pages 3 – 13, 2011