Automatic recognition of plasma anomaly behaviours in massive databases: potential application to ITER

J. Vega¹, R. Castro¹, S. Dormido-Canto², G. A. Rattá¹, M. Ruiz³

¹Laboratorio Nacional de Fusión, CIEMAT, Madrid, Spain
²Dpto. Informática y Automática - UNED, Madrid, Spain
³Instrumentation and Applied Acoustic Research Group, UPM, Campus Sur, Madrid, Spain
Motivation

- Nowadays, processing all information of a fusion database is a much more important issue than acquiring data
  - Massive databases
- Fusion devices produce tens of thousands of discharges but only a very limited part of the collected information is analysed
  - Physics studies normally limited to a few tens of shots
- Plasma behaviours are recognised in experimental signals by the identification of known patterns
  - Diagnostics produce the same morphological patterns in the signals for reproducible plasma behaviours
- The analysis of physical events requires their identification and temporal location
  - The recognition and location of events are the main concerns in relation to the analysis: manual, complex and very time consuming searching processes
- Long pulse devices (W7X or ITER) will have databases with very large number of signals and very long records
  - ITER: discharges 30 minutes long and up to $10^6$ signals per discharge
Motivation

• Can we identify relevant temporal segments in an automatic way?
  • ‘relevant’ means ‘with interest from some point of view’: either physics or machine control

• An automatic first screening of discharges would allow focusing the analysis on a reduced set of time intervals
  • Irrelevant parts of the shots are discarded
  • Improvement of statistical relevance for known events
    • Practically all the information inside the databases can be used
  • Potential detection of unknown events that appear on a regular basis

• Automatic recognition of relevant temporal segments means
  • Reduction of human efforts
  • Standardisation of criteria
    • It reduces the vulnerability to human errors: missing occurrences, subjective assessments or location errors
Outline

• Recognition of relevant events

• Algorithm to identify relevant events by detecting anomalies in signals

• Specific methods to detect anomalies in signals

• Conclusions
How to proceed?

• Big Data techniques deal with heterogeneous, complex and massive datasets to identify patterns that are hidden inside enormous volumes of data

• ITER is expected to acquire more than 1 TB of data per discharge

• Signals can be time/amplitude series, temporal evolution of profiles (amplitude/radius relationship) and video-movies (infra-red and visible cameras)

• W7X or ITER databases satisfy the conditions of heterogeneity, complexity, size and hidden patterns to use Big Data techniques
How to recognise relevant events?

- A relevant event can be any kind of perturbation in the plasma evolution
- This is revealed in the temporal evolution of signals by means of unexpected variations (anomalies)
  - Time series
    - Amplitude, noise, presence/suppression of patterns with periodical structure
  - Profiles
    - Amplitude, hollow profiles, peaked profiles, wider profiles, gradients
  - Video-movies
    - Emission increasing
- An automatic search for events will have to locate anomalies in individual signals
- Interesting plasma behaviours are usually recognised by simultaneous anomalies in several signals
To perform automatic recognition, software codes have to be executed in an unattended way

- 1\textsuperscript{st} step: to define a dataset of signals and a range of discharges
Algorithm for off-line automatic recognition of relevant events: 6 step process

To perform automatic recognition, software codes have to be executed in an unattended way

- 1\textsuperscript{st} step: to define a dataset of signals and a range of discharges
- 2\textsuperscript{nd} step: to determine times in each discharge when individual signals show anomalies

\begin{itemize}
  \item Anomaly times in each signal of a discharge
  \item 4 anomalies in this case
\end{itemize}
Algorithm for off-line automatic recognition of relevant events: 6 step process

To perform automatic recognition, software codes have to be executed in an unattended way

• 1\textsuperscript{st} step: to define a dataset of signals and a range of discharges
• 2\textsuperscript{nd} step: to determine times in each discharge where individual signals show anomalies
• 3\textsuperscript{rd} step: to chose the morphological patterns within a time interval $t_S$ around the anomaly time
Algorithm for off-line automatic recognition of relevant events: 6 step process

- 3rd step: to choose the morphological patterns within a time interval $t_S$ around the anomaly time
  - The time interval $t_S$ corresponding to the same plasma event could be quite different in several occurrences (in the same shot or in different shots)
  - The definition of the time interval means two selections: the starting time and the temporal length $t_S$
- How to decide the interval of the different signals in an unattended way?

4 anomalies in this example => 4 potential relevant events

Dataset
- $N_S$ signals/discharge
- $N_D$ discharges

\[
\text{if } A_j, j = 1, ..., N_D \text{ is the number of anomalies in shot } j
\]

Total number of potential relevant events: $A_{\text{TOTAL}} = \sum_{j=1}^{N_D} A_j$
Algorithm for off-line automatic recognition of relevant events: 6 step process

To perform automatic recognition, software codes have to be executed in an unattended way

- 1\textsuperscript{st} step: to define a dataset of signals and a range of discharges
- 2\textsuperscript{nd} step: to determine times in each discharge where individual signals show anomalies
- 3\textsuperscript{rd} step: to chose the morphological pattern of each individual signal within a time interval $t_S$ around the anomaly time
- 4\textsuperscript{th} step: to define multi-signal patterns (MSP)
Algorithm for off-line automatic recognition of relevant events: 6 step process

• 4th step: to define multi-signal patterns (MSP)
  • A MSP is made up of all patterns of all signals determined in step 3 around a common anomaly time
    • Signals without recognition of anomaly are also part of the MSP
  • A MSP is characterised by the morphological patterns of all the signals with a common time interval
  • A criterion to define the common time interval is necessary taking into account all MSPs in all discharges of the dataset
    • All MSPs need to have the same dimensionality

By assuming 200 sampling times per MSP and $N_S = 100$ signals with
  • 95 time series
  • 3 profiles (120 points each)
  • 2 video-movies (500x300 each) and 2 bytes per sample, the total amount of memory is 120 Mbytes/MSP
Algorithm for off-line automatic recognition of relevant events: 6 step process

To perform automatic recognition, software codes have to be executed in an unattended way

• 1\textsuperscript{st} step: to define a dataset of signals and a range of discharges
• 2\textsuperscript{nd} step: to determine times in each discharge where individual signals show anomalies
• 3\textsuperscript{rd} step: to chose the morphological pattern of each individual signal within a time interval $t_S$ around the anomaly time
• 4\textsuperscript{th} step: to define multi-signal patterns (MSP)
• 5\textsuperscript{th} step: to group the MSPs into a number of sensible clusters in an unsupervised way (this reveals the organisation of the MSPs)
Algorithm for off-line automatic recognition of relevant events: 6 step process

• 5\textsuperscript{th} step: to group the MSPs into a number of sensible clusters in an unsupervised way (this reveals the organisation of the MSPs)
  • The grouping of the MSPs into clusters provides the classification of the relevant events
  • The different clusters can be labelled but the challenge is to identify each cluster with a physical behaviour of the plasma
    • To be done by experts NOT in unattended way
  • Clusters that are identified with physical behaviours can be used to increase the statistical relevance of the data analysis
  • Clusters that are not identified with physical behaviours but show statistical weight suggest the presence of plasma behaviours not recognised so far
  • Clusters without statistical weight can be considered outliers

• By assuming 1 relevant event/10 s the unsupervised classification process requires 720 Mbytes/minute per shot
• Thinking of ITER shots (30 minutes long), this implies 21 Gbytes of memory per shot
• By considering $N_D = 500$ discharges, the total memory amount to solve the unsupervised clustering is 10 Tbytes!!
  • The curse of dimensionality
• High performance computing is needed
Algorithm for off-line automatic recognition of relevant events: 6 step process

To perform automatic recognition, software codes have to be executed in an unattended way

• 1\textsuperscript{st} step: to define a dataset of signals and a range of discharges
• 2\textsuperscript{nd} step: to determine times in each discharge where individual signals show anomalies
• 3\textsuperscript{rd} step: to chose the morphological pattern of each individual signal within a time interval $t_s$ around the anomaly time
• 4\textsuperscript{th} step: to define multi-signal patterns (MSP)
• 5\textsuperscript{th} step: to group the MSPs into a number of sensible clusters in an unsupervised way (this reveals the organisation of the MSPs)
• 6\textsuperscript{th} step: to develop supervised classifiers with the classes of step 5
  • Classification of new MSPs with confidence measures allows assessing the reliability of the whole process
  • In this step, classes of MSPs are well-defined
  • Supervised classifiers can be implemented under real-time conditions
How to determine anomalies in individual signals?

• Anomalies in the temporal evolution of signals translate the existence of changes in the plasma behaviour
  • The more abrupt the change of shape in a signal the more abrupt the change in the plasma evolution

• Our analysis has been based on recognising changes in individual signals
  • This allows establishing the potential set of signals related to each plasma behaviour

• Each anomaly has to include a time interval around its time value
  • The objective is to try the characterisation of the several plasma behaviours by combining the several shapes of the signals around the anomalies

• Methods to locate anomalies in signals should provide an estimation of the time interval around the anomaly
Methods to determine anomaly times and related time interval: 1

- Detection of outliers through a generalised linear regression model
  - If the temporal evolution is smooth, amplitudes between consecutive samples are very similar
  - In a space $Y(t - \tau) - Y(t)$, samples are distributed along the diagonal
  - Samples outside the diagonal are outliers
    - These are identified as outliers in the normal probability plots of residuals

- The number of consecutive samples that are outliers determine the time interval
  - If the sampling period is $\tau$, in both cases the time interval is $10 \cdot \tau$
Methods to determine anomaly times and related time interval: 2

- **Use of martingales for testing exchangeability**
  - The only assumption in the data stream is the *iid* hypothesis
    - Samples are independent and identically distributed (*iid*)
  - Anomalies are detected as the samples are produced
  - Anomalies are recognised when the martingale crosses the $\lambda$ threshold
  - The assumed rate of false alarms is $1/\lambda$

![Graph showing p-values follow U(0, 1)](image1)

- The time interval of the anomaly corresponds to the time in which the martingale increases to achieve the $\lambda$ value

![Graph showing lambda = 10](image2)
Related works

Anomaly detections

• Universal Multi-Event Locator (UMEL)

• Spectral clustering optimization

• To follow the temporal evolution of the Fourier components of a signal
  • R. Castro et al. 12th IAEA Control, DAQ & RP. May 13-17, 2019

• Using deep learning methods
  • G. Farias et al. 12th IAEA Control, DAQ & RP. May 13-17, 2019
  • G. Farias et al. O-31 (Friday)

Real-time implementations

• Methodologies to standardise the development of FPGA-based intelligent DAQ systems on heterogeneous platforms using OpenCL
  • M. Astrain et al. 12th IAEA Control, DAQ & RP. May 13-17, 2019
Conclusions

• Big data techniques will be essential for the automatic location and classification of plasma anomaly behaviours

• Methods for the automatic discovering of anomalies in signals have been discussed

• An algorithm to relate multi-signal patterns in an automatic way has been established

• Unsupervised classifications will allow labelling the clusters
  • High performance computing is needed
  • The correspondence between labels and physics behaviours has to be decided by experts

• Unsupervised clusters can be converted into reliable supervised classifiers
  • Real-time applications are possible
Thank you very much for your attention!