EFDA–JET–CP(08)06-03

A. Pereira, J. Vega, A. Portas, R. Castro, A. Murari
and JET EFDA contributors

Optimized Search Strategies to Improve Structual Pattern Recognition Techniques

Optimized Search Strategies to Improve Structual Pattern Recognition Techniques

A. Pereira[1], J. Vega[1], A. Portas[1], R. Castro[1], A. Murari[2]
and JET EFDA contributors*

JET-EFDA, Culham Science Centre, OX14 3DB, Abingdon, UK

[1]Asociación EURATOM/CIEMAT, Avda. Complutense 22, 28040 Madrid, Spain
[2]Consorzio RFX-Associazione EURATOM ENEA per la Fusione. I-35127 Padua, Italy
** See annex of M.L. Watkins et al, "Overview of JET Results ",*
(Proc. 21 st IAEA Fusion Energy Conference, Chengdu, China (2006)).

Preprint of Paper to be submitted for publication in Proceedings of the
8th International FLINS Conference on Computer Intelligence in Decision and Control, Madrid, Spain.
(21st September 2008 - 24th September 2008)

ABSTRACT.
Structural pattern recognition techniques are an efficient way to apply a pattern oriented data retrieval paradigm. Some techniques have already been implemented in the JET Analysis Cluster (JAC) by means of a general purpose tool (software application) to allow the identification of similar patterns (structural shapes) inside temporal evolution signals. Data retrieval methods are based on three essential aspects: feature extraction (to reduce signal dimensionality), the classification system (to index objects according to some criteria) and similarity measure (to compare how similar two objects are), but there is not a single solution or unique criterion to handle these key elements. This paper provides a new solution to the localization and extraction of similar patterns in time-series data. Alternative searches are proposed to objectively increase the recognition of similar patterns so as to achieve better results on the data retrieval. In the proposed approach, patterns are represented by string of characters. Looking for patterns means looking for characters. The recognition problem is translated into a character-matching problem. Thinner search strategies have been studied with excellent results in the detection of long subpatterns. Long subpatterns are not so easy to identify since even a single mismatch in one character can compromise similarity between two patterns. Identifying long patterns in a fast, fault tolerant and intelligent way is the aim of the analyzed strategies, formally based on statistical criteria and some aspects of probability theory.

1. INTRODUCTION

Some pattern recognition methods for data retrieval have already been applied to fusion databases. The first approach was focused on looking for similar full waveforms [1] (the shape of an entire signal) and later, the interest was concentrated on searching for specific patterns within waveforms, developing two different techniques for this task (based on equal length segments [2] and variable length segments [3]). A segment is the result of applying the least square minimization procedure to obtain a fitted straight line. The slope of each segment is classified according to a discrete set of values (alphabet code) that can be stored as a sequence of characters in a relational database, allowing the use of well formed query sentences to search for specific patterns. A software application for data retrieval (including entire waveform search and pattern search within signals) is available in a concurrent way from the JAC Linux cluster at JET.

In general, the feature extraction and the classification systems have obtained good results in the detection of general subpatterns. With the extended use of the application, it has been observed that long subpatterns are not always matched correctly. Studying this problem, we have detected that the number of encoded segments and the number of utilized primitives (the selected set of characters) are essential elements to obtain good results. In addition, a new primitive computation is presented to simplify the waveform encoding, to increase the search efficiency and to optimize the data retrieval for long patterns.

2. OPTIMIZING THE PATTERN RECOGNITION TECHNIQUES

The methods described in this paper have been focused on optimizing the already existing techniques.

As in a preceding approach, the original signal is decomposed into equal line segments. However, in this case, each segment is the result of applying a delta transformation to consecutive Haar wavelet coefficients of the entire signal [4] (δ is the difference between sequential data). Then the slope of a segment is $\delta/\Delta x$. If we apply the Haar transform,Δx is constant because it represents the same time period. Thus, Δx is the unique value that we need to know. The entire time discharge in JET pulses is of the order of 40s and reducing a signal to 64 Haar coefficients, we get a resolution of $\Delta x = 625$ mseg. Therefore, 64 Haar coefficients are an adequate value to decompose any signal without large losses of information. After this processing step, the segments are labeled according to the sign value of the delta transformation, i.e. positive (primitive a), negative (primitive b). The feature vector (set of primitives) and the delta values for each signal are the quantities stored in the relational database.

The selection of only 2 primitives is the essential difference with regard to previous developments. By using just two primitives, the probability that the longest pattern (the entire waveform) will happen is $1/2^{64}$ (2 primitives and 64 delta values) very much higher than in the previous encoding $1/4^{128}$ (4 slopes and 128 constant segments). With only two primitives and 64 delta values, we have increased the possibility to match longer subpatterns hidden in large amount of data.

3. SIMILARITY QUERIES

Given a query Q (pattern to seek) and N objects (stored signals in the database) where everyone has M (64 in this case) data sequences (set of primitives), we can do either exact searches (whole matching) or approximate searches (partial matching). Exact searches consist of locating data subsequences inside the objects that match a query sequence exactly (complete query). Approximate searches locate data subsequences that match a query subsequence (incomplete query). Finally, to identify a subsequence that is most similar to the one of interest, a similarity measure (the Euclidian distance) is defined to be able to compare how similar two subsequences are. This query range allows finding every object with the better distance from the query Q. If only exact searches are performed, short subsequences will be found with higher probability (excess of results) but long subsequences will be found with a lower statistically probability (lack of results) due to very small and sometimes irrelevant mismatch with respect to the reference pattern. As these sequences become longer, the probability of matching should be increased by allowing more flexible queries (approximate similarity). This would permit to match many sequences with high relationship independently of the pattern length. To avoid recovering too short sequences, it is possible to demand that the pattern length has a dimension or minimal size.

4. IMPROVING THE DATA RETRIEVAL

Delta values in Nuclear Fusion data often represent a small change near zero, either positive or negative. Figure 1 shows the frequency distribution of the whole set of deltas typical of Electron Cycloctron Emission (ECE) signals used to measure the electron temperature of the plasma. This has been fitted with a normal probability density function with parameters given by the sample

mean 142.85 and standard deviation $\sigma = 38960$ of the data.

The likelihood function (1), is the normal probability density function at each of the x values (deltas):

$$y = f(x|\mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \; e^{\frac{-(x-\;)^2}{2\sigma^2}} \qquad (1)$$

Almost all delta values are grouped in the middle of the distribution and due to the fact that very small jumps near zero can adopt both values (positive or negative) without scarcely any physical significance (these variations can be due simply to noise), we can compose powerful queries fitting these primitives with any value, increasing in this way, the probability of matching long subpatterns in the relational database. The flexibility level required in the query will depend on which proportion of cases falls into the indistinct central category.

A statistical analysis is the best way to determine the range of the δ variations where the primitives can take any value.

According to our analysis of the ECE signals, the choice $[-\sigma, \sigma]$ (Figure 2), is not a good selection because there will be a lot of indifferent values (high number of cases that fall into the central category, indistinctly 'a' and 'b', near to zero), and only a little bit of significant primitives, i.e. 'a' or 'b'. To increase the number of suitable primitives it is necessary to reduce the interval to a range of values where they do not contain useful information (close to zero). This was achieved choosing the range $[-\sigma/32, \sigma/32]$ (Figure 3). Thus, the probability of indifferent values has been reduced and the suitable primitives (far to zero) appear with higher frequency. With this feature we compose powerful queries, getting a large relationship of similar patterns in the data retrieval.

CONCLUSIONS

The probability to match many patterns has been increased allowing better scores than previous results. Partial matching is an approximate way to amplify the similarity between longer subpatterns and whole matching is an exact way to search the shortest subpatterns. Hence, the possibility to locate a lot of similar subsequences is less dependent on the pattern length. In addition, statistical criteria were studied to identify a stable decision on the primitive assignment that can be suitable to any kind of signal and also no dependent of the database size. These intelligent strategies have been added to the former similarity algorithms and have been successfully implemented on the tool already available in the JAC Linux cluster at JET.

REFERENCES

[1]. J. Vega, A. Pereira, et al., *"Data mining technique for fast retrieval of similar waveforms in Fusion massive databases"*. Fusion Engineering and Design. Vol **83**, Issue 1, pp 132-139. January 2008

[2]. Dormido-Canto S., Farias G., et al., *"Search and retrieval of plasma waveforms: structural pattern recognition approach"*. Rev. Sci. Ins. **77** (2006). 10F514

[3]. Vega J., Rattà G., et al., *"Recent results on structural pattern recognition for Fusion massive databases"*. Proc. of the IEEE International Symposium on Intelligent Signal Processing. ISBN: 1-4244-0829-6 (2007) 949-954

[4]. Vega J., Murari A., et al., *"Structural Pattern Recognition Techniques for Data Retrieval in Fusion Massive Databases"*. International Workshop on Burning Plasma Diagnostics. Villa Monastero, Varenna, Italy. 24-28 September (2007).

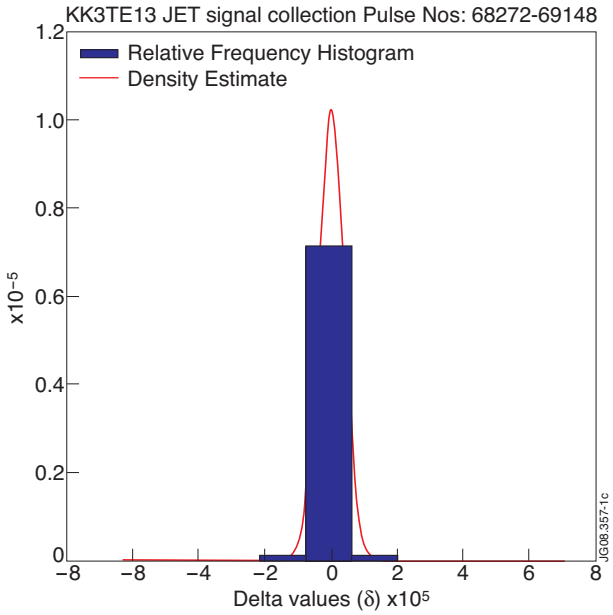

Figure 1. Frequency distribution of the whole set of deltas and the normal probability density function for an ECE signal at JET.

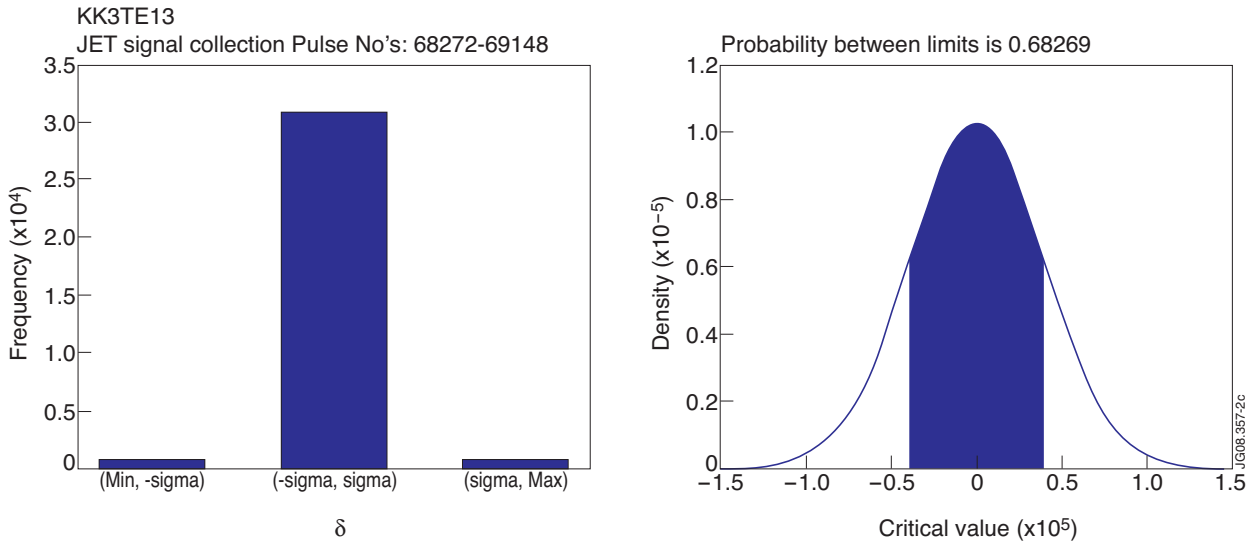

Figure 2: If we choose negative sigma and positive sigma as arbitrary central boundaries, the probability to find a delta value between these central limits is the area closed between the curve and the lower and upper limits.

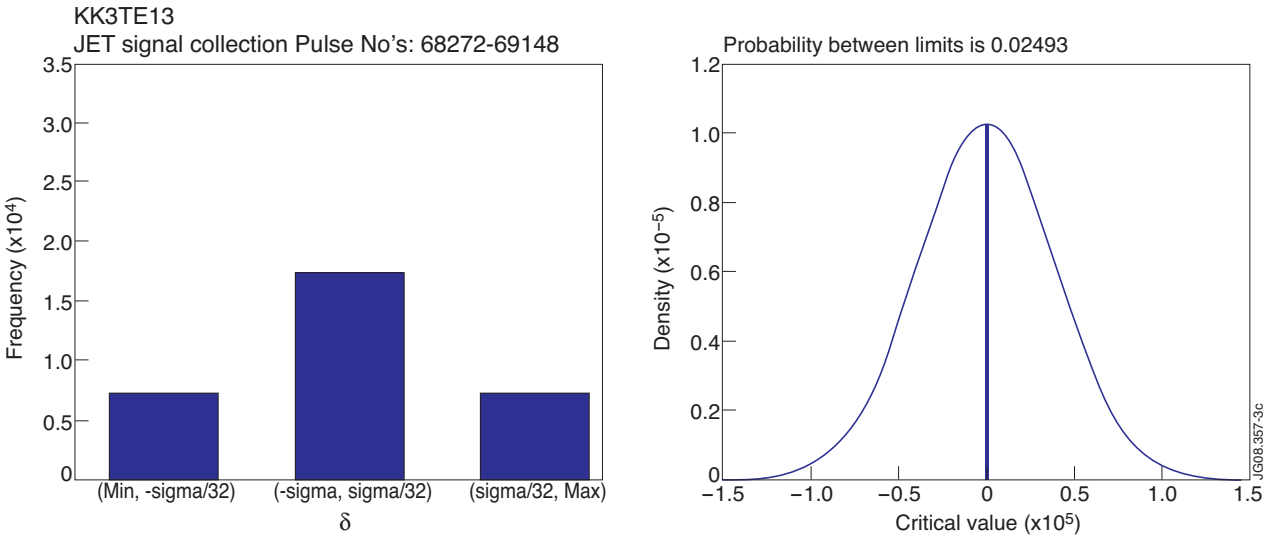

Figure 3: Frequency histogram and probability with [-√/32, √/32] as central boundaries.