Harmonics Systems for Time Mining

Daniela López De Luise[1]
*(CIIS Lab, Buenos Aires, Argentina)

ABSTRACT: *Information management is currently one of the topics with most impact. Processes are studied in this paper in their temporality, mainly to be able to make predictions. Interesting information arises when what is being examined is not the complex information produced along the development of a process, but the temporality itself for the process. Such an alternative point of view enables the relation of this analysis to a real process, and its application in the very moment the events are taking place, knowing that there is a starting point of certain a priori knowledge.*

This work is shows this alternative stance, from the temporal perspective. There are applications of concepts and analogies brought from a different environment where time is handled in a more natural way. We shall call this approach "Harmonic Systems". We shall present the basis for harmonic systems, the focused approach we propose as a type of time mining, and its application.

Keywords: *Data Mining, Information Systems, Process Mining, Process Modeling, Soft Computing*

I. INTRODUCTION

Information management is currently one of the topics with most impact. It relates to a broad range of matters from those related to engineering processes to others such as speech processing and consumer profile generation. It is possible to obtain information with special types of mining. For example, there are mining techniques in the area of language that allow for a relatively efficient management of automated translation between different languages [1][2], detection of a writing style in an author, the study of the effects of bilingualism, etc. [3][4][5][6]

From the standpoint of mining, processes are studied in their temporality, mainly to be able to make predictions (of times, events and characteristics) and descriptions (of use, of information type, business profiles, user profiles, etc). [7] [8] [9]

Techniques are usually crucial for figuring out which information is concurrent or subsequent to a certain fact [10] [11]. But an interesting aspect arises when what is being examined is not the complex information produced along the development of a process, but the temporality itself for the process. Taking an opposite stance to the traditional one and eschewing the study of the characteristics of the process in favor of its rhythm, accelerations, static periods and any other aspect that can be related to the measurement of time features. Such an alternative view enables the relation of this analysis to a real process, and its application in the very moment the events are taking place, knowing that there is a starting point of certain a priori knowledge.

The main aim of this work is to show this alternative stance, from the temporal, through an application of concepts and analogies brought from a different environment where time is handled in a more natural way. We shall call this approach "Harmonic Systems". In this work we shall present the basis for harmonic systems (Section 2), the focused approach we propose as a type of time mining (Section 3) and a test application demo (Section 4).

II. HARMONIC SYSTEMS

2.1. Harmonic Systems Purpose

The production of meta-data from information generated in a sequence of data associated with a time evolution. These meta-data are subject to change with time (as opposed to those models created through mining) and comprise a model of dynamic approach with a certain behavior that typically evolves in time, while keeping its own identity pattern. Thus, a natural process could result in one of more patterns with the same or a different time variation. On the other hand, with the application of filters, a selection can be made of a certain time sub-sequence inside said temporal variation.

2.2.Harmonic Systems Application

Generating the model is just an intermediate step that defines whether a sequence resulting from a pattern is being produced. This way, it is possible to test the patterns of interest that could belong to software or hardware faults, concurrent processes, etc., while the information is being produced.

In these cases, the objective is to be able to act in the moment where the change is happening, with no regard to the temporal variations in its evolution, but focusing instead on the pattern. The search for accuracy is out of the question, as the aim is to detect the appearance of the pattern with a definite degree of certainty.

A comparison could be drawn to an analogous situation of a doctor evaluating a patient in given moment of the evolution of its illness: if he shows a certain sustained temperature, congestion and presence of nasal secretion, there could be further symptoms, but a decongestant is needed.

2.3. General Characteristics of Harmonic Systems

There are different types of harmonics. An harmonic is a combination of properties that are of interest, which in this context are referred to as "pattern". They represent one or more current subsets. Example pattern 1: ID-pattern=1, threshold U=0.3

T	Property-1	Property-2
t1=λ_1	PROC=A	USR=034
t2=λ_2	PROC=C	USR=035
t3=λ_3	PROC=A	USR=035

Table 1. Pattern 1

It can be observed that the pattern is a sequence of three moments t_1, t_2, t_3, which must happen one after the other to be complete. We shall see below how a pattern can trigger actions in $1<t<3$ according to its associated triggering threshold. This concept will be explained further along.

An harmonic is in resonance when the sequences of changes in the values of the properties inside the pattern are in correspondence with those in the dataset that is being processed.

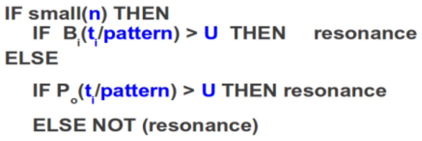
When an harmonic begins to resonate, a series of processes is triggered, being the monitoring and updating of the process the most important for the model. This is achieved simply by capturing the current sequence and adding it to the model.

2.4. Resonance

It comprises the event of detection of compatibility of an harmonic with certain data vectors. This comparison consists of two steps:

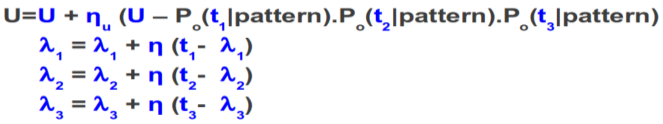
a) Pattern detection: patterns are evaluated according to the properties that describe them. When there is a matching with current data (in Example 1: Property-1=A, Property-2=034), the resonance must be checked, comparing the probability of the pattern against the threshold U (0.3 for the example).

b) Resonance: In case the resonance is verified, the model is updated in the following manner:
-Action P, which is associated to the patter, is triggered to produce meta-data and tracking data.
-The temporality of each step is captured (t_1, t_2 and t_3 in the example) as a difference between them.
-The size of **n** is compared against a certain cut-off threshold n_c (e.g., n_c=80). When $n<n_c$, small(n) is true, otherwise it is false. When small(n) gives true, the Binomial dispersion of harmonics is assumed, otherwise it is considered to be Poisson:



```
IF small(n) THEN
    IF $B_i(t_i/pattern) > U$ THEN resonance
ELSE
    IF $P_o(t_i/pattern) > U$ THEN resonance
    ELSE NOT (resonance)
```

Then U_i is updated and the new parameters with a Hebbian type learning [12], assuming the use of Poisson:



$$U=U+\eta_u\,(U-P_o(t_1|pattern).P_o(t_2|pattern).P_o(t_3|pattern)$$
$$\lambda_1=\lambda_1+\eta\,(t_1-\lambda_1)$$
$$\lambda_2=\lambda_2+\eta\,(t_2-\lambda_2)$$
$$\lambda_3=\lambda_3+\eta\,(t_3-\lambda_3)$$

where h is a rate of learning for the parameters, a weighing factor for the relative importance of the pattern fulfillment. At the same time, the new threshold U will reflect variations with a sensitivity given by $\mathbf{h_u}$. The values for **h** and $\mathbf{h_u}$ do not need to be the same.

2.5. Types of Harmonics

-Harmonics are data vectors with a threshold of tolerance to difference. Inside that threshold, data is considered to be resonating.

Characteristic 1: time data is used as a difference in respect to the previous time, $t=(t_i - t_{i-1})$, as opposed to the classic techniques [13] where the value of a certain property is compared in a time t in respect to t_{i-1}, and the measured magnitude or effect becomes a property associated to this time differential.

Consequently: there is no comparison of series length nor corrections in them due to a difference in length. There is also no normalization.

Characteristic 2: given that no component alignment is required between patterns, distance has no need for corrective techniques [13] such as Dynamic Time Warping, Longest Common Subsequence Similarity, local Scaling Functions, global scaling function, etc.

Characteristic 3: the temporal mining approach takes the properties being measured as a ***pattern*** which identifies the components in a time series, and models the time dispersion instead of the set of properties inside the pattern.

This calls for the development of a model of the pattern, one which could be analogous to Probabilistic similarity measure where methods are ***model*** based. They provide the ability to incorporate prior knowledge into the similarity measure. However, it is not clear whether other problems such as time-series indexing, retrieval and clustering can be

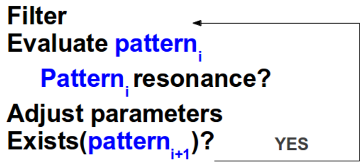
efficiently, general similarity approach involving a transformation rules language [4], and hundreds of algorithms from Data Mining to classify, cluster, segment and index time series.
-Data will have a degree of belonging to the harmonic it resonates with, which would decrease with the distance of the current t_i values in relation to those expected.

2.7.Filter Types

Data can previously go through some of the filters described below. The effect is focusing only in critical harmonics for the purposes of this study.
There are three filter types:
-High-pass filters: They leave the data that are beyond a certain distance (δ) that $t_i + \delta < t_{actual}|pattern$. Since the pattern's property $p_1..p_i$ is met, t_i exceeds the model value).
-Low-pass filters: They leave the data that are closer than a certain distance ($t_i - \delta > t_{actual}|pattern$**).** Since the pattern's property $p_1..p_i$ is met, t_i is lower than the model value).
-Band-pass filters: They leave the data that are within a certain distance range ($t_i + \delta > t_{actual}|pattern > t_i - \delta$). Since the pattern's property $p_1..p_i$ is met, t_i is within the model value with a certain distance).

2.8.Operating Mechanism

The data set is vectorized and then used based on a global distance. All filters will use this distance in their activities. Database vectors go through the following cycle:



Note that in each cycle the model of the pattern that feeds the subsequent cycle is optimized, therefore, if the model is stored, it is possible to star with a better description of the data –supposing fluctuation between starts is not wider.

III. TIME AS TEMPORALITY MINING

3.1. Temporality Mining Purpose

Performing mining based on merely temporal processes as a mechanism to assess and predict events that are concomitant with them. Among other things it is possible to obtain information regarding the typical evolution of these events, the importance of the event within the life of the system being studied.

3.2. Temporality Mining Characteristics

Data are considered if they meet the following requirements:
-They are captured from a process which is external to the system
-They have marked temporality: date, time, minutes, seconds, etc.
-They can be associated to a cyclic complex event or not, thus enabling generation of an approximate model starting from a measurable characteristic set and its temporality.
-They are produced by one or more identifiable and distinguishable source
-Events are defined and limited in time
-Their duration is variable or constant but they start and end at a defined time point
-They may undergo duration and frequency changes, but not changes in identifying characteristics –properties defined as behavior patterns.

3.3. Data Formats suitable for Temporality Mining

A temporality mining vector in this context would be something like :

Header + characteristics + body

with Header= timestamp,
Characteristics= properties to be assessed taken to a numeric vector of numeric labels or not, which do not belong in a metric scale. E.g.: characteristics= process-ID+param$_1$+param$_2$+param$_3$, with param$_1$, param$_2$, param$_3$ =1 if these parameters are provided; otherwise, 0
Body= other dataset information not relevant to the current process

3. 4.Applications of Temporality Mining

It is possible to apply this type of mining to different situations, mainly prediction of time variations, detection/removal of noise sources (data which do not fit in the time pattern expected), detection of mutations in temporal characteristics and detection of cycles and periodicities and their characteristics. There are many applications, including user profiling, intrusion detection, fault prediction, operation profile, product/buyer association profile, etc.

IV. HARMONICS AND TEMPORALITY MINING

The process explained in the temporality mining section is applied, adapting it to the special case of time mining with data format:

HEADER + CHARACTERISTICS + BODY + PROCESS

Where:

HEADER= timestamp

CHARACTERISTICS= numeric vectors with the values of properties to analyze.

E.g.:

characteristics= process-ID+param$_1$+param$_2$+param$_3$, (with param$_1$, param$_2$, param$_3$ =1 if these parameters are provided; otherwise, 0)

Body= other dataset information not relevant to the current process

PROCESS= the process triggered when current characteristics are met

4.1. Testing

The dataset was taken from [14]. Raw data is something as:

123.123.123.123 - - [26/Apr/2000:00:23:47-0400] "GET /asctorf/http/1.0" 200 8130 "http://search.netscape.com/.../RTF" "Mozilla/4.05 (Macintosh; I; PPC)", which includes fields like IP, used ID, timestamp, result code, etc.

Data was processed and changed to the following format:

{head} + {characteristics} + {body}

For temoprality mining purposes it is evident that the heading must be a time related field, such as the timestamp. Characteristics hold for properties of special interest during the current study:

{timestamp} + {ID-process, IP} + {rest of the information in the row}

From the given example the information entries result as:

20000426002347 +{GET, 123.123.123.123} + {/asctorf/http/1.0 200 8130 "http://search.netscape.com/.../RTF" "Mozilla/4.05 (Macintosh; I; PPC)}

4.2. Test 1: Temporality with harmonics

For demonstrative purposes, a test with 14 samples was performed. For all cases, the commands were replaced with numeric identifiers. E.g.: GET is replaced by 1.

The pattern to be studied was defined as:

pattern= 1,
U=0.03,
n_+ = 33,
n_T=70,
□=0.05,
□$_u$=0.3,
n_c=56

and IP=123.123.123.123, l_1=20.7, l_2=28.5, and l_3=67.5, n=14

Poisson is used to measure the pattern model against the patterns obtained from the data and confirm whether they resonate.

In the processing, no filter was applied, and for pattern detection the steps are as follows.

In ID =4 the first part of the following pattern is detected:

t2=* - 20000426002393,
t1=20000426002393- 20000426002372= 21

In ID=6 the second part of the pattern is detected:

t3=* - 20000426002423,
t2= 20000426002423 - 20000426002393= 30,
t1=20000426002393- 20000426002372= 21

In ID=9 the third part of the pattern is detected:

In ID=12, the pattern is completed –simplified for practical purposes, supposing the start of another activity finished the previous one:

$$
\begin{aligned}
&t3= 20000426002483 - 20000426002423=60,\\
&t2= 20000426002423 - 20000426002393= 30,\\
&t1=20000426002393- 20000426002372= 21
\end{aligned}
$$

For resonance verification: Resonance to the only pattern in process is verified, taking the Cumulative Poisson Distribution:

$$
\begin{aligned}
&P_o(17.5, 21) \text{ x } P_o (22.5, 30)\text{x } P_o (67.5, 60) :: U=0.03\\
&0.585 \text{ x } 0.658 \text{ x } 0.198 :: 0.03\\
&0.076 :: 0.03
\end{aligned}
$$

Since aggregated Poisson production was the threshold U, the pattern resonates and is considered compatible with the calculated model.

When the pattern is met, the model is updated, considering this case as favorable.

Since parameters are just an average of the number of tics, the average of tics is updated based on the new values found:

$$
\begin{aligned}
&\lambda_1=20.7 + 0.05*(21-20.7) = 20.72\\
&\lambda_2=28.5 + 0.05* (30-28.5)= 28.60\\
&\lambda_3=67.5 + 0.05 * (60-67)= 63.15
\end{aligned}
$$

The pattern threshold is also updated:

$$
\begin{aligned}
&U= U+ \eta_u \text{ x } (U - P_o(17.5, 21) \text{ x } P_o (22.5, 30)\text{x } Po (67.5, 60))\\
&U= 0.03+0.3*(0.03 - (0.585 \text{ x } 0.658 \text{ x } 0.198))= 0.03 + 0.3 * (0.076 – 0.03) = 0.04
\end{aligned}
$$

Temporality mining with harmonics without threshold change:

$$t1=*- 20000426002372$$

It is possible to use the thresholds in a fixed way, in which case the information is collected precisely when the pattern is met as expected, and there are no attempts to model actual processes, but instead the purpose is to detect the patterns that fit into a given model (maybe to separate them).

4.3. Test 2: Using Filters

A high-pass filter was defined with a threshold of 60 time units and the process was previously performed with this filter. The system administrator intends to detect processes which take longer than usual.

The pattern to be studied was defined as: **pattern-ID= 2, U=0.053, n_+ = 34, n_T=71**, **n_c=80,** with initial parameters l_1=20.72, l_2=28.60, and l_3=63.15

Since $\mathbf{n_T}$ is not larger than $\mathbf{n_c}$ samples, Binomial distribution is used to measure the pattern model against the patterns obtained from the data and confirm whether they resonate. Due to space reasons, the dataset is not detailed here. For information purposes, the dataset comprises 996 samples selected with processes 1, 2 and 3 and enabled IPs were: *123.123.123.123*, *123.123.123.120*, *123.123.123.119* and *123.123.123.123.110*

As preliminary step, the high-pass filter is applied, removing 58 of the 996 samples (5.8%) The remaining samples go through the process. During pattern detection 11 pattern occurrences were detected. It is worth noting that, if the above filter had not been applied, the patterns detected would have been 13. These 11 patterns correspond to 33 of a total of 354 inputs within the logs generated by the IP *123.123.123.123*

Resonance was verified for each instance applying the formulas given in section I, item D (resonance). As instances were processed, the parameters were adjusted. U_f started to stabilize towards a higher value, represented the closeness of current processes to the model. At the same time, l_i values evolved from their initial values.

It is remarkable that almost all the patterns survived the high-pass filter, if $\mathbf{n_c}$ is the maximum delay, it could point to a problem of long times in the processes involved. From this point of view, it is confirmed by the parameter values obtained: l_1=21.13, l_2=28.10, l_3=64.26. Both process 1 and 3 should be revised since they are expected to be faster.

V. CONCLUSION AND FUTURE WORK

This work presented the *Harmonic Systems* model for mining in real time. The strategy is focused only in times and behavioral patterns. This approach may be a useful for studies related to certain automated processes and monitor hardware systems.

Other interesting alternatives should be analyzed and filter cadences need to be defined in a more sophisticated way (e.g., Gaussian). Also, alternative methods for the initial model should be specified and their efficacy compared. Reasonable alternatives to define the value of parameter δ should be analyzed.

REFERENCES

[1] C. Faloutsos, M. Ranganathan, Y. Manolopoulos, Fast subsequence Matching in Time SeriesDatabases. *Proc ACM SIGMOID. 1994*

[2] R. Agrawal, K. Lin, H. Sawhney, K. Shim, *Fast similarity search in the presence of noise, Scaling and translation in time series Databases. Proc VLDB 95. pp 490-501, 1995*

[3] D. López De Luise, A Morphosyntactical Complementary Structure for Searching and Browsing, *In Proc. of SCSS05. Springer. 2005*

[4] D. López De Luise, MLW and bilingualism. *Adv. Research and Trends in New Technologies, Software, Human-Computer Interaction, and Communicability. IGI Global. USA. 2013*

[5] K. Church, R. Mercer, Introduction to the Special Issue on Computational Linguistics Using Large Corpora. *Computational Linguistics. MIT Press Cambridge, MA, USA . 1993*

[6] A. Kao, S. Poteet, Natural Language Processing and Text Mining (*Springler. 1991)*

[7] F. Tak-chung, Engineering Applications of Artificial Intelligence. *Engineering Applications of Artificial Intelligence 24, pp. 164–181. 2011*

[8] J. Han, M. Kamber, Mining Stream, Time-Series, and Sequence Data.*In Data Mining, Second Edition, Concepts and Techniques. 2nd Edition. 2011*

[9] J. Shieh, E. Keogh, iSAX: Indexing and Mining Terabyte Sized Time Series. *Proceedings KDD'08. pp 623-631. ACM. 2008*

[10] H. Jagadish, A. Mendelzon, Similarity based queries. *In Proc 14th PODS 95. pp 36-45. 1995*

[11] B. Bollogás., G. Das, D. Gunopulos, H. Mannila, Time series similarity problems and well-separed geometric sets. *Nordic Journal of Computing. Pp 409- 423. 2003*

[12] L. Clifford, Neural Networks. *Teoretical Foundations and Analysis (IEEE Press.1991)*

[13] C.A. Ratanamahatana, J. Lin, D. Gunopulos, E. Keogh. Mining Time Series Data. *In Data Mining and Knowledge Discovery Handbook, pp 1049-1077. 2010.*

[14] Apache logs: http://www.herongyang.com/Windows/Web-Log-File-IIS-Apache-Sample.html