QSI - Qualitative Similarity Index

F.J. Cuberos
Dpto. Planificación-Radio Televisión de Andalucía
Ctra. San Juan - Tomares Km. 1.3.
San Juan de Aznalfarache. Sevilla
fjcuberos@rtva.es

J.A. Ortega and **R.M. Gasca** and **M. Toro**
Departamento de Lenguajes y Sistemas Informáticos.
University of Seville
Avda. Reina Mercedes s/n. Sevilla
{ortega, gasca, mtoro}@lsi.us.es

Abstract

There are different approaches to the temporal study of time evolving systems. In this paper, this study is carried out by means of the comparison of time series.
It is proposed as an improvement in the comparison of time series with the inclusion of qualitative knowledge. Taking into account the evolution of the values of the series, our approach uses a similarity index defined by qualitative labels. Every label represents a rank of values that we may consider similar, from a qualitative perspective. The proposed index is defined by means of the matching of qualitative labels.
Let be a time series, a label is obtained with every transaction of every two adjacent values. This label depends on the magnitude and the sign of the transaction. If every label is represented by means of a single character, then the evolution of the temporal series is translated into a string. Finally, an index of similarity of the time series is defined according to the similarity of the obtained strings.
This proposed index has been applied to the dataset of Australian signs (Australian Sing Language Dataset) of UCI KDD with a correct identification rate superior to the 95 per cent. In this paper, it has been applied to study the different behaviours of a semiqualitative model of logistic growth with a delay.

Introduction

The study of temporal evolution of systems is an incipient research area. It is necessary the development of new methodologies to analyze and to process the time series obtained from the evolution of those systems. These time series are usually stored in databases. It is necessary to develop new algorithms for its study.

A time series is a sequence of real values, each one represents the value of a magnitude at a point of time. A possible field of application is the comparison of time series in numeric databases. We are interested in databases obtained from the evolution of dynamic systems. It is proposed in (Ortega *et al.* 99) a methodology to simulate semiqualitative dynamic systems. These simulations are stored into a database. This database may also be obtained by means of the data acquire from sensors installed in the real system. There are a variety of applications to produce and to store time series.

When we are working with time-series databases, one of the biggest problems is to calculate the similarity between two given time series. The interest of a similarity measure is multiple. In this paper, this interest is focused on: finding the different behaviour patterns of the system stored in a database, looking for a particular pattern, reducing the number of relevance series before applying analysis algorithms, etc.

Assuming that the similarity is a distance function of the time series, we catalogue the basic queries to manage a time series database in three groups:

• *Range query:* given a series, finding those series that are similar in within a distance.

• *Nearest neighbor*: given a series, finding in the database the series which is the nearest neighbor in accordance with a defined distance.

• *All-pairs query*: finding all the pairs of series in the database that are within a distance of each other.

Many approaches have been proposed to solve the problem of an efficient comparison. In this paper, we proposed to carry out this comparison from a qualitative perspective, taking into account the variations of the time series values. The idea of our proposal is to abstract the numerical values of the time series and to concentrate the comparison in the shape of the time series.

In this paper, we do not take into account time series with noise, it is postponed for future work.

The rest of this paper is structured as follows: first, we analyze some related works that we have used to define our index. Next the *Shape Definition Language* is introduced, which is appropriate to carry out the translation of the original values, and we also explain the problem of the *Longest Common Subsequence* (LCS). Next section introduces our approach, the *Qualitative Similarity Index*. Finally, this index is applied to a semiqualitative logistics growth model with a delay.

Related Work

In the literature, different approximations have been developed to study time series. In (Agrawal *et al.* 95b) present the shape definition language (SDL). (SDL), which is suitable for retrieving objects based on shapes contained in the his-

tories associated with these objects. An important feature of the language is its ability to perform blurry matching where the user cares only about the overall shape. This work is the key to translate the original data into a qualitative description of its evolution that allows a subsequent comparison.

On the other hand, those works that study the problem of the Longest Common Subsequence (LCS) are also related to this paper, because we use (LCS) algorithms as the baseline to define our index. (Paterson&Dancík94) collect a complete review of most known solutions to this problem.

There has been many works on comparison of time series (Faloutsos *et al.* 94). Most of them propose the definition of indexes, which are applied to a subset of values extracted from the original data. These indexes provide an efficient comparison of time series. They are defined taking into account only some of the original values. This improvement of speed produces a decrease in the accuracy of the comparison.

These indexes are obtained applying a transformation from the time series values to a lower dimensionality space. Other approaches differ in the way to carry out this mapping or in the selected target space.

One option is to select only a few coefficients of a transformation process to represent all the information of the original series. In this approach, we find the change from the time domain to frequency domain. In (Agrawal *et al.* 95a), it is used the Discrete Fourier Transform (DFT) to reduce the series to the first Fourier Coefficients. In (Chan&Waichee99), it is proposed a solution based in the Discrete Wavelet Transform (DWT) in a similar way.

Other approaches reduce the original data in the time series, selecting a subset of the original values. In (Keogh&Pazzani98), it uses a piece-wise linear segmentation of the original curve. In (Keogh&Pazzani99), the Dynamic Time Warping (DTW) algorithm is applied over the segmented data, and finally in the work (Keogh&Pazzani00) it is made a straight dimensionality reduction with Piecewise Constant Approximation, selecting a fixed number of values of the original data. It is known as *PCA-indexing*.

The last option is to generate a 4-tuple-feature vector extracted from every sequence. In (Kim *et al.* 01), this vector is proposed and a new distance function is defined as the similarity index.

In the paper (Cheung&Stephanopoulos90), it is proposed the study of series with different time scales from a qualitative perspective.

Shape Definition Language (SDL)

This language proposed in (Agrawal *et al.* 95b) is very suitable to create queries about the evolution of values or magnitudes along the time.

For any set of values stored for a time period, the fundamental idea in SDL is to divide the range of the possible variations between adjacent values in a collection of disjoint ranges and to assign a label for each of them.

Figure 1 represents a sample division in three regions of the positive axis. This division depends on the possible variations and the assigned labels. The behaviour of a series

Figure 1: Possible assignment of labels

may be described taking into account the transitions between consecutive values. A derivative series is obtained by means of the difference of amplitude among the consecutive values of the time series. The value of this difference match in one of the disjoint ranges, and therefore this value so defined produces a label of the alphabet.

This translation produces a transitions sequence based on an alphabet. The symbols of this alphabet describe the magnitude of the increments of the values of the time series. Every symbol is defined by means of four descriptors. The firsts two are the lower and upper bounds of the allowed variation from the initial value to the final value of the transition. The last two specify the constraints on the initial and final value of the transition, respectively.

The alphabet proposed in (Agrawal *et al.* 95b) has only 8 symbols. The size of this alphabet is very small compared with the real values that the series may achieve along time. This translation prioritized the shape over the original values of the time series. This affirmation will be later explained. Figure 2 shows an example of translation using the set of symbols ($Down, down, stable, zero, up, Up$).

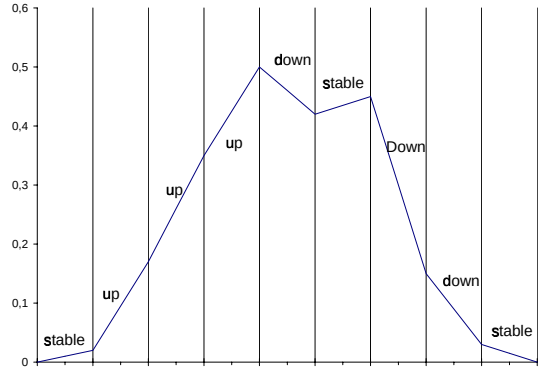

Figure 2: Example of translation

Every string of symbols may describe an infinite number of curves. All of them verify the constraints imposed for the symbols to the represented transitions.

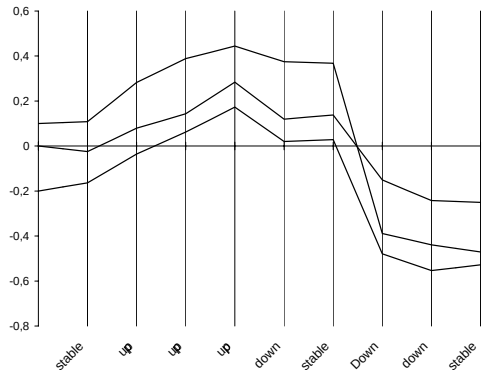

Figure 3: Translation with identical sequence

Figure 3 shows three different curves with the same sequence of symbols, even though the curves have different initial points. In a similar way, we use a language to translate a time series described by means of their numerical values into a string of symbols which represent the variations between adjacent values.

Longest Common Subsequence (LCS)

Working with different kinds of sequences, from strings to DNA chains, one of the most used similarity measure is the *Longest Common Subsequence* (LCS) of two or more given sequences. LCS is the longest collection of elements appears in both sequences and in the same order. For two finite sequences $s = \langle s_1, s_2, ..., s_m \rangle$,$t = \langle t_1, t_2, ..., t_n \rangle$ over alphabet $\sum$, the LCS problem consists of finding the longest sequence $\langle u_1, u_2, ..., u_r \rangle$ such that there exist indices $i_1 < i_2 ... < i_r$, with $1 \leq r \leq m$, and $j_1 < j_2 ... < j_r$, with $1 \leq r \leq n$, such that $s_{i_z} = u_z$ and $t_{j_z} = u_z$.

The algorithms to compute LCS are well known and a deeper analysis is not necessary. Different algorithms for LCS has been analyzed and a deeper analysis compared in (Paterson&Dancík94).

Our interest in LCS come from:

• The SDL language generates a string of symbols from the original numeric values of the time series, so it is possible to apply the LCS algorithm to find a "distance" between two time series, abstracting the shapes of the curves.

• The LCS is a special case of the Dynamic Time Warping (DTW) algorithm reducing the distance increment of each comparison to 0 or 1 depending on the presence, or not, of the same symbols. So LCS inherits all the features of DTW.

DTW is an algorithm intensively used in speech recognition area because it is appropriate to detect similar shapes that are non aligned in the time axis. This lack of alignment induces catastrophic errors in the comparison of shapes which use the Euclidean distance.

The idea of DTW is to find a set of ordered mappings between the values of two series, so the global distance o warping cost is minimized.

Qualitative Similarity Index (QSI)

The idea of this index is the inclusion of qualitative knowledge in the comparison of time series. It is proposed a measure based in the matching of qualitative labels that represent the evolution of the series values. Each label represents a range of values that may be assumed as similar from a qualitative perspective . Different series with a qualitatively similar evolution produce the same sequence of labels.

The proposed approximation performs better comparisons than previously proposed methods. This improvement is mainly due to two characteristics of the index: it maximizes the exactness because it is defined using all the information of the time series; and on the other hand, it focuses the comparison on the shape and not on the original values because it considers the evolution of groups as similar. It is interesting to note that we suppose that the time series are noise free between to samples and with a linear and monotonic evolution.

Let $X = \langle x_0, ..., x_f \rangle$ be a time series. Our proposed approach is applied in three steps. First, a normalization of the values of X is performed, yielding $\tilde{X} = \langle \tilde{x_0}, ..., \tilde{x_f} \rangle$. Using this series we obtain the differences series $X_D = \langle d_0, ..., d_{f-1} \rangle$, that it is translated to a string $S_X = \langle c_1, ..., c_{f-1} \rangle$. The similarity between two time series is calculated by means of the comparison of the two strings obtained from them, applying the previous transformation process, and then using the LCS algorithm. The result is used as a similarity measure between the original time series.

Normalization

Keeping in mind the qualitative comparison of the series, it is made a normalization of the original numerical values in the interval [0,1]. This normalization is carried out to allow the comparison of time series with different quantitative scales.

Let $X = \langle x_0, ..., x_f \rangle$ be a time series, and let $\tilde{X} = \langle \tilde{x_0}, ..., \tilde{x_f} \rangle$ be the normalized temporal series obtained from X, as follows:

$$\tilde{x_i} = \frac{x_i - min(x_0, ..., x_f)}{max(x_0, ..., x_f) - min(x_0, ..., x_f)} \quad (1)$$

where min and max are operations that return the maximum a minimum values of a numerical sequence, respectively.

Let $X_D = \langle d_0, ..., d_{f-1} \rangle$ be the series of differences obtained from $\tilde{X}$ as follows:

$$d_i = \tilde{x_i} - \tilde{x}_{i-1} \quad (2)$$

This difference series will be used in the labeling step to produce the string of characters corresponding to X. It is interesting to note that every $d_i \in X_D$ is a value in the [-1,1] interval, as a consequence of the normalization process.

Labeling process

The proposed normalization in the previous section is focused in the slope evolution and not in the original values. A label may be assigned to every different slope, so the range

of all the possible slopes is divided into groups and a qualitative label is assigned to every group.

The range division is defined depending on the parameter δ which is supplied by the experts according to their knowledge about the system. The value of this parameter has a direct influence in the quality of the results, therefore this is an open research area of this paper that we will detail in future work.

Label	Range	Symbol
High increase	$[1/\delta, +\infty]$	H
Medium increase	$[1/\delta^2, 1/\delta]$	M
Low increase	$[0, 1/\delta^2]$	L
No variation	0	0
Low decrease	$[-1/\delta^2, 0]$	l
Medium decrease	$[-1/\delta, -1/\delta^2]$	m
High decrease	$[-\infty, -1/\delta]$	h

Where the first column represents the qualitative label for every range of derivatives, which is shown in the second row. Last column contains the character assigned to each label. The proposed alphabet contains three characters for increases and three for decreases ranges, and one additional character for constant range. It is important to note that in our approach there is no application of the constraints presented in SDL (Agrawal *et al.* 95b).

This alphabet is used to obtain the string of characters $S_X = \langle c_1, ..., c_{f-1} \rangle$ corresponding to the time series X, where every c_i represents the evolution of the curve between two adjacent time points in X and it is obtained from $X_D = \langle d_0, ..., d_{f-1} \rangle$ assigning to every d_i its character in accordance with the above table.

This translation of the time series to a sequence of symbols lets us abstract from the real values and focus our attention on the shape of the curve. Every sequence of symbols describes a complete family of curves with a similar evolution.

Figure 4 shows a normalized curve with their derivative values and the assigned label to each transition between adjacent values. This example has been obtained selecting $\delta = 5$.

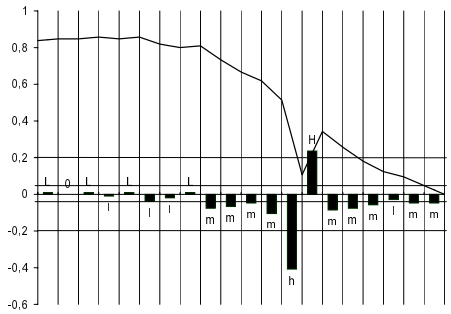

Figure 4: Sample of translation

Definition of QSI Similarity

Let X, Y be time series where $X = \langle x_0, ..., x_f \rangle$ and $Y = \langle y_0, ..., y_f \rangle$. Let S_X, S_Y be the strings obtained when X, Y are normalized and labelled.

The QSI similarity between the strings S_X, S_Y is defined as follows

$$QSI(S_X, S_Y) = \frac{\nabla(LCS(S_X, S_Y))}{m} \tag{3}$$

where ∇S is the counter quantifier applied to string S. Counter quantifier yields the number of characters of S. On the other hand, m is defined as $m = max(\nabla S_X, \nabla S_Y)$. Therefore, the QSI similarity may be understood like the number of ordered symbols that we may find in the same order in both sequences simultaneously, and this value divided by the length of the longest sequence.

Properties of QSI We are going to described two properties of QSI.

Let S_X, S_Y, S_Z be three strings of characters obtained from the transformation of three temporal series X, Y, Z respectively. The definition of QSI similarity verifies that:
Property 1. $QSI(S_X, S_Y)$ is a number in the interval $[0, 1]$. If X is absolutely different of Y it is 0.

$$\begin{array}{c} \text{If } LCS(S_X, S_Y)) = \emptyset \\ \Downarrow \\ \nabla LCS(S_X, S_Y)) = 0 \\ \Downarrow \\ QSI(S_X, S_Y) = 0. \end{array} \tag{4}$$

The $QSI(S_X, S_Y)$ value increases according to the number of coincident characters. This number is 1 if $S_X = S_Y$.

$$\begin{array}{c} \text{If } LCS(S_X, S_Y)) = S_X \\ \Downarrow \\ \nabla LCS(S_X, S_Y)) = S_X \\ \Downarrow \\ QSI(S_X, S_Y) = 1. \end{array} \tag{5}$$

Property 2. The length of the strings to compare also has an important influence. In this sense, two strings with approximate lengths and with a number of coincident symbols are more similar than two strings with the same number of coincident symbols but with different lengths:

$$\begin{array}{c} \nabla S_X \approx \nabla S_Y, \nabla S_X \approx \nabla S_Z, \\ \nabla LCS(S_X, S_Y) \approx \nabla LCS(S_X, S_Z) \\ \Downarrow \\ QSI(S_X, S_Y) > QSI(S_X, S_Z) \end{array} \tag{6}$$

Comparison with other approach

When a new approach is introduced, it is interesting to test its validity and the improvement with respect to the other approaches that appeared in the literature. In this paper, we are going to compare our approach with the algorithm introduced in (Keogh&Pazzani99), called Segmented Dynamic Time Warping ($SDTW$). (Keogh&Pazzani99) carries out a clustering process with a set of time series. Every clustering process joins sets of data in subsets trying the similarity among the elements of every subset to be minimized and the similarity between different subsets to be minimized too.

The $SDTW$ algorithm was tested with the Australian Sign Language Dataset from the UCI KDD (Bay99) choosing 5 samples for each word. The data in the database are the 3-D position of the hand of five signers, records by means of a data glove.

In order to carry out the comparison between both approaches, we have chosen the same 10 words used in (Keogh&Pazzani99) from the 95 words included in the database. Next, for every possible pairing of different words (45), we have clustered the 10 sequences (5 of each word), using a hierarchical clustering using the average, with two different distance measures. First, we used the distance defined in the classic DTW algorithm applied to the original numerical values of the series. The result was 22 correct clustering from 45. Next, we used the similarity QSI index, proposed in this paper, over the string obtained from the translation of the original values of the series. This time, the result was 44 correct clustering of 45.

The total success obtained with DWT is exactly the same reported by (Keogh&Pazzani99), but the success obtained with QSI similarity is better.

Application of QSI to a logistics growth model with a delay

The following generic names: logistic, sigmoidal, and s-shaped processes are given to those systems in which an initial phase of exponential growth is followed by another phase of approaching to a saturation value asymptotically (figure 5). This growth is exhibited by those systems for which exponential expansion is truncated by the limitation of the resources required for this growth.

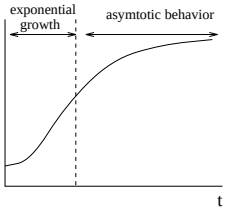

Figure 5: Logistics growth curve

In literature, these models have been profusely studied. They abound both in natural processes, and in social and socio-technical systems. These models appear in the evolution of bacteria, in mineral extraction, in world population growth, economic development, learning curves, some diffusion phenomena within a given population such as epidemics or rumors, etc. In all these cases, their common behaviours are shown in figure 6. There is a bimodal behaviour pattern attractor: A stands for normal growth, and O for decay. It can be observed how it combines exponential with asymptotic growth. This phenomenon was first modelled by the Belgian sociologist Verhulst in relation with human population growth in (Verhulst50). Nowadays, it has a wide variety of applications, and some of them have just been mentioned.

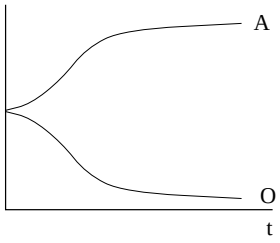

Figure 6: Logistics growth model

Let S be the qualitative model. If we add a delay in the feedback paths of S, then its differential equations are

$$\Phi \equiv \begin{cases} \dot{x} = x(n\,r - m), \\ y = delay_\tau(x), \quad x > 0, \quad r = h_1(y), \\ h_1 \equiv \{(-\infty,-\infty),+,(d_0,0),+,(0,1), \\ \qquad +,(d_1,e_0),-,(1,0),-(+\infty,-\infty)\} \end{cases}$$

being n the increasing factor, m the decreasing factor, and h_1 a qualitative continuous function defined by means of points and the derivative sign among two consecutive points. These functions are explained in detail in (Ortega 2000). This function has a maximum point at (x_1, y_0). The initial conditions are

$$\Phi_0 \equiv \begin{cases} x_0 \in [LP_x, MP_x], \\ LP_x(m), \\ LP_x(n), \\ \tau \in [MP_\tau, VP_\tau] \end{cases}$$

where LP, MP, VP are the qualitative unary operators *slightly positive, moderately positive and very positive* for the x, τ variables.

The methodology described in (Ortega *et al.* 99) is applied to this model to obtain the database of time series. This methodology transforms this semiqualitative model into a family of quantitative models. Stochastic techniques are applied to choose a quantitative model of the family. The simulation of every selected quantitative model generates a time series that is stored into the database. We would like to classify the different behaviours of the system applying the QSI similarity to the obtained database. Figure 7 contains the table obtained when this index is applied between every two time series of the database. In this figure some of these time series are shown .

Similarity matrix obtained with *QSI*

54X	55X	1X	18X	77X	17X	73X	Series
0,87	0,872	0,41	0,44	0,494	0,43	0,376	50X
	0,994	0,292	0,314	0,388	0,384	0,35	54X
		0,294	0,316	0,39	0,384	0,35	55X
			0,758	0,792	0,598	0,586	1X
				0,754	0,58	0,552	18X
					0,632	0,62	77X
						0,93	17X

Figure 7: QSI similarity of the model

Three different behaviours in this table appears according to the obtained value QSI, they have been remarked in figure 8.

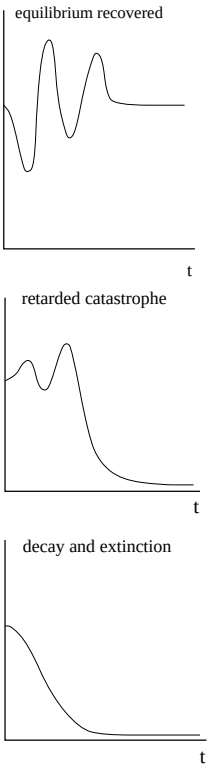

Figure 8: Logistics growth model with a delay

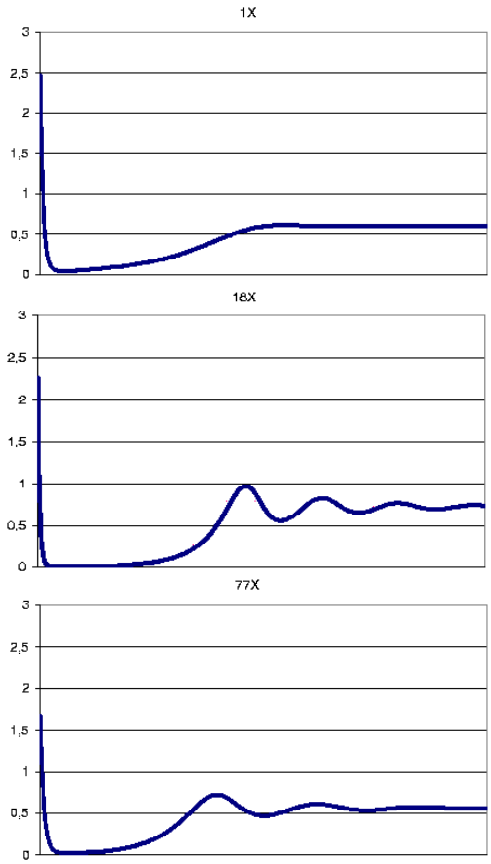

Figure 9: Recovered equilibrium

The results obtained in this way to discover the behaviour patterns are in accordance to others appeared in the bibliography (Aracil *et al.* 97) and (Karsky *et al.* 1992) where the results are concluded by means of a mathematical reasoning.

Figures 9, 10 and 11 show the time series grouped by these behaviours. Figure 9 shows the 1x,18x and 77x time series whose behaviour is classified as *recovered equilibrium*. In a similar way, figure 10 shows the 17x and 73x time series whose behaviour is labelled as *retarded catastrophe* and finally, in figure 11 is shown the 50x, 54x and 55x time series whose behaviour is labelled as *decay and extinction*.

Conclusions and Further Work

In this paper, we have introduced the QSI index to measure the similarity of time series depending on its qualitative features. Furthermore the proposed method achieves better results than previous algorithms with a similar computational cost.

In order to apply the QSI similarity index between time series, it is necessary a normalization process of the time series. Next, the sequence of differences is obtained from this normalized series . Finally, this sequence is translated into a string using a qualitatively defined alphabet. The LCS algorithms are used to calculate the index QSI. The results obtained are in accordance with other previous works, although our approach produces a better classification.

In the future, the idea is the automation and the optimization of the division in ranges of the possible slopes by studying the number of regions and its limits.

Other further works are directed to study the application of this technique to time series with noise, and to study the possibility to define similarity grades, applying the index to models with different time scales.

Acknowledgments

This work was partially supported by the Spanish Interministerial Committee of Science and Technology by means of the programs DPI2001-4404-E and DPI2000-0666-C02-02.

References

Agrawal R., Lin K.I., Sawhney H.S. and Shim K. Fast similarity search in the presence of noise, scaling, and translation in time series databases. *The* 21^{st} *VLDB Conference* Switzerland, (1995).

Agrawal R., Psaila G., Wimmers E.L. and Zaït M. Querying shapes of Histories. *The* 21^{st} *VLDB Conference* Switzerland, pp. 502-514 (1995).

Aracil J., Ponce E. and Pizarro L. Behavior patterns of logistic models with a delay *Mathematics and computer in simulation* 44: 123–141, (1997).

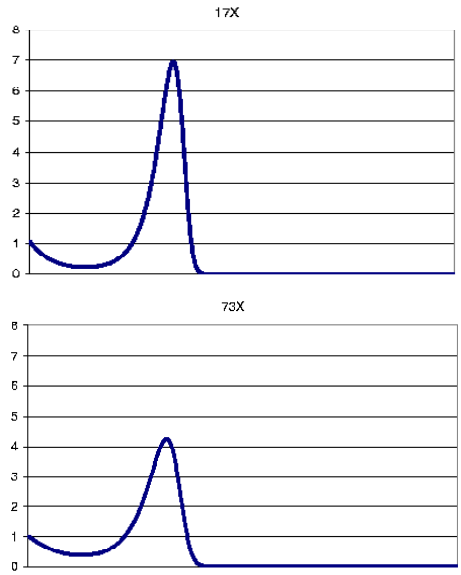

Figure 10: Retarded catastrophe

Bay S. UCI Repository of KDD databases (http://kdd.ics.uci.edu/). Irvine, CA: University of California, Departamet of Information and Computer Science. (1999).

Chan K. and Wai-chee F.A. Efficient time series matching by wavelets *Proc.* 15^{th} *International Conference on Data Engineering*, (1999).

Cheung J.T. and Stephanopoulos G. Representation of process trend - Part II. The problem of scale and qualitative scaling, *Computers and Chemical Engineering* 14(4/5), pp. 511-539, (1990).

Faloutsos C., Ranganathan M., and Manolopoulos Y. Fast subsequence matching in time-series databases. *The ACM SIGMOD Conference on Management of Data*, pp. 419-429 (1994).

Karsky M. Dore J.-C. and Gueneau P. Da la possibilité d'apparition de catastrophes différès. *Ecodecision No 6*, (1992).

Keogh E.J. and Pazzani M.J. An enhaced representation of time series wich allows fast and accurate classification, clustering and relevance feedback *Proc.* 4^{th} *International Conference of Knowledge Discovery and Data Mining*, pp. 239-241, AAAI Press (1998).

Keogh E.J. and Pazzani M.J. Scaling up Dynamic Time Warping to massive datasets, *Proc. Principles and Practice of Knowledge Discovery in Databases*, (1999).

Keogh E.J. and Pazzani M.J. A simple dimensionality reduction technique for fast similarity search in large time series databases, (2000).

Kim S-W, Park S. and Chu W.W. An Index-Based Approach for Similarity Search Supporting Time Warping in Large Sequence Databases. *Proc. 17th IEEE Int'l Conf. on Data Engineering*, Heidelberg, Germany, (2001).

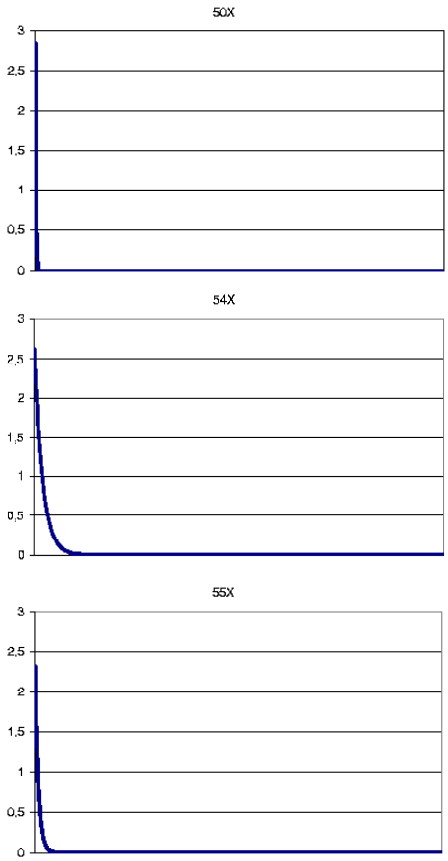

Figure 11: Decay and extinction

Ortega J.A., Gasca R.M. and Toro M. A semiqualitative methodology for reasoning about dynamic systems. 13^{th} *International Workshop on Qualitative Reasoning*. Loch Awe (Scotland), 169–177, 1999.

*Ortega J.A.*Patrones de comportamiento temporal en modelos semicualitativos con restricciones. Ph.D. diss., Dept. of Computer Science, Seville Univ, (2000).

Paterson M. and Dancík V. Longest Common Subsequences. *Mathematical Foundations of Computer Science* vol. 841 de LNCS, pp.127-142, (1994).

Verhulst P. F. A Quetelet, *Annuaire de l'Académie royale des sciences de Belgique* 16 , pp.97-124, (1850).