Computación y Sistemas Vol. 5 No. 2 pp. 96 - 108
© 2001, CIC - IPN. *ISSN 1405-5546* Impreso en México

Comparación Cualitativa de Series Temporales. Índice Cualitativo de Similitud - QSI

Qualitative Comparison of Time Series. Qualitative Similarity Index - QSI

Juan Antonio Ortega[1], Francisco Javier Cuberos[2], Rafael M. Gasca[1] y Miguel Toro[1]
[1]Departamento de Lenguajes y Sistemas Informáticos, Universidad de Sevilla
Avda. Reina Mercedes s/n, Sevilla
[2]Departamento de Planificación, Radio Televisión Andalucía
Carretera San Juan - Tomares Km. 1.3 San Juan de Aznalfarache, Sevilla
e-mail : {ortega, gasca, mtoro}@lsi.us.es, fjcuberos@rtva.es

Artículo recibido en mayo 30, 2001; aceptado en septiembre 25, 2001

Resumen

En este documento proponemos el estudio temporal de sistemas que evolucionan en el tiempo mediante la comparación de series temporales. Continuando el trabajo (Ortega et al., 2000) se propone una mejora en la forma de comparar series temporales con la incorporación de conocimiento cualitativo, mediante etiquetas cualitativas. Cada etiqueta representa un rango de valores que, desde una perspectiva cualitativa, podemos considerar similares. La elección de etiquetas de un solo carácter nos permite la aplicación de algoritmos de comparación cadenas. Finalmente definimos un índice de similitud de las series basado en la similitud de las cadenas en que han sido convertidas.

Palabras clave: Series Temporales, Conocimiento Cualitativo, Modelos Semicualitativos.

Abstract

In this paper we propose the study of systems that evolve in the time by means of the comparison of temporary series. Continuing the work (Ortega et al., 2000), we carry out an improvement in the form to compare temporal series with the incorporation of qualitative knowledge by means of qualitative labels. Each label represents a rank of values that, from a qualitative perspective, we can consider similars. The selection of labels of a single character allows us the application of algorithms of string comparison. Finally we defined an index of similarity of the series based on the similarity of the strings which they have been translated.

Keywords: Temporary Series, Qualitative Knowledge, Semiqualitative Models.

1 Introducción

El estudio de los sistemas que evolucionan en el tiempo es un área de investigación muy actual. Es necesario desarrollar metodologías que permitan analizar las series temporales obtenidas de la evolución de estos sistemas. Normalmente, se generan bases de datos y es necesario desarrollar algoritmos para su análisis con la intención de identificar concordancias entre los datos.

Una serie temporal es una secuencia de números reales, cada número representando el valor de una magnitud en un instante de tiempo. Un posible campo de aplicación de estas metodologías es comparar series temporales en bases de datos numéricas. En nuestro caso estamos interesados en bases de datos obtenidas como resultado de la evolución de sistemas dinámicos. Estas bases pueden ser generadas de diferentes formas: en (Ortega *et al.*, 1999) se propone una metodología para obtener una base de datos y otra posibilidad es obtenerlas mediante el registro de los datos proporcionados por sensores instalados en el sistema. En cualquier caso son muy numerosas las aplicaciones prácticas que generan y almacenan series temporales.

Uno de los mayores problemas en el uso de bases de datos de series temporales es la similitud entre dos series cualesquiera. El interés en una medida de similitud es múltiple: obtener diferentes patrones en la base de datos, buscar uno determinado, reducir la base con aquellas series que sean similares con la intención de que los algoritmos que se aplican para el análisis de la base actúen sólo sobre los datos relevantes, etc.

Considerando que la similitud venga determinada por

una función de distancia entre las serie, podemos catalogar las diferentes consultas básicas para la manipulación de una base de datos de series temporales en tres tipos:

• Consulta de rango: donde se desean localizar todas las series que se distancian en menos de un ϵ indicado de una serie determinada.

• Vecino más próximo: Dada una serie ha de localizarse aquella que está mas cerca a ella que ninguna otra de la base.

•Pares cercanos: Con esta consulta se desea encontrar todos los pares de series presentes en la base de datos que está a una distancia menor que un ϵ indicado.

Además de esto podemos estar interesados en la localización de subsecuencias dentro secuencias completas.

Se han realizado múltiples aproximaciones al problema de la comparación eficiente pero en este trabajo presentamos la posibilidad de que esta comparación se realice desde un punto de vista cualitativo sobre las variaciones de los valores de las series. Con esta metodología se pretende abstraer los valores concretos de la serie a la forma de la curva que los representa.

En el presente trabajo no se consideran las bases de datos con ruido, cuyo estudio y aplicación se postpone para futuros trabajos.

El resto de este documento se estructura como sigue. En la siguiente sección se hace una revisión de los trabajos relacionados, luego presentamos el Lenguaje de Definición de Formas en el que se basa la traducción de los valores originales, y seguidamente se describe el problema de la obtención de la Subcadena Común Máxima (LCS). En la siguiente sección se introduce nuestra propuesta, esto es, el índice cualitativo de similitud, Finalmente, se aplica este índice al estudio de un modelo semicualitativo de crecimiento logístico con retraso.

2 Trabajos Relacionados

Existen diferentes alternativas a la hora de comparar las series de datos.

En (Agrawal *et al.*, 1995b) se presenta un lenguaje de definición de Formas (SDL) para recuperar curvas de una base histórica basándose en la forma de las curvas. Una importante característica de este lenguaje es su capacidad para hacer búsquedas difusas donde el usuario sólo indica la forma general de lo que busca por medio de descripciones realizadas con el propio lenguaje. Este trabajo es fundamental para convertir los datos originales en una descripción cualitativa de su evolución que nos permita después una comparación.

El presente trabajo se basa en los estudios existentes sobre el problema de la subsecuencia común máxima (LCS) al utilizar los algoritmos conocidos como base de la definición de nuestro índice. En (Apostolico, 1997) se realiza una revisión de muchas de las soluciones existentes.

Por otro lado, en la bibliografía se han presentado diferentes aproximaciones para comparar series temporales. La mayoría proponen la creación de un índice con un pequeño conjunto de valores extraídos de los datos originales.

Los índices presentados en estos trabajos proporcionan una comparación eficiente de las series temporales consiguiendo una velocidad de comparación muy superior a la obtenida con la comparación de todos los datos originales.

Para la generación del índice se ha optado mayoritariamente por dos enfoques diferentes: el realizar una transformación de los valores de la serie temporal a un espacio de menor dimensión y la de reducir directamente los datos originales de la serie temporal seleccionando un subconjunto de ellos.

Vamos a dedicar una subsección a cada uno de éstos grupos de trabajos y una final a otras opciones.

2.1 Utilización de Transformadas

Los diferentes trabajos difieren en la manera de hacer esta transformación o en el espacio sobre el que se realiza. Esta aproximación se basa en la idea de que las series, entendidas como ondas,caracterizados por tener una gran concentración de energía en unos pocos armónicos.

Cuando estamos hablando de series de datos sería más correcto referirnos a que lo que se produce es una concentración de la información.

En (Agrawal *et al.*, 1993) se propone la utilización de la Transformada Discreta de Fourier (DFT) para reducir la serie a sus primeros coeficientes de Fourier. Apoyándose en el teorema de Parseval y en los resultados de (Oppenheim *et al.*, 1975).

La transformación se basa en que el paso desde el dominio del tiempo al de la frecuencia deja invariable la

distancia entre series.

Una vez reducidas las series a los k principales coeficientes DFT, éstos se consideran como un punto en un espacio característico de $2k$ dimensiones con el que se construye un índice denominado *índice F*.

Con estos puntos se genera el índice basado en árboles $R*$ (Beckmann *et al.*, 1990) y se utiliza la distancia euclídea como función de distancia. Así para realizar una consulta en la que se indica una secuencia Q y una tolerancia ϵ el procedimiento es: obtener los coeficientes DFT de la secuencia Q proporcionado un punto q_f; usar el índice para obtener los puntos con una distancia menor que ϵ a q_f; filtrar las falsas alarmas.

Se demuestra que este enfoque no desecha ninguna serie válida al realizar una consulta, pero proporciona un conjunto ampliado de las respuestas correctas que debe ser procesado para desechar estos casos no válidos.

Abundando en la misma dirección que el trabajo anterior y utilizando el mismo mecanismo de indexación en (Rafiei and Mendelzon, 1998) se obtiene una aceleración con un factor de 2 utilizando, además de los k primeros coeficientes DFT, los k últimos. Los últimos coeficientes sólo se utilizan en el cálculo de la distancia y no forman parte del índice al verificarse que son los complejos conjugados de sus correspondientes homólogos de la lista de coeficientes.

Ampliando el trabajo de (Agrawal *et al.*, 1993) en (Rafiei and Mendelzon, 1997) se utiliza el segundo y tercer coeficiente DFT junto con los datos descriptores de la forma normal de cada serie, según (Goldin and Kanellakis, 1995), para obtener un índice almacenado en un árbol R*. Pero a diferencia de aquél, en que el sentido de la similitud indicada por un cierto valor ϵ está fijado por la definición del algoritmo, éste define una clase de transformaciones de desplazamiento, reescalado y media móvil que permiten definir al usuario el tipo de similitud que se desea buscar.

Por su parte, en (Rafiei 1999) se permite que el concepto de similitud sea definido en función de múltiples transformaciones dando de esta manera una mayor libertad al usuario para realizar las búsquedas de elementos similares. Para ello se presenta un nuevo algoritmo de búsqueda que permite aplicar la colección de transformaciones de forma simultánea en un único recorrido del índice.

En (Faloutsos *et al.*, 1994) se amplía el trabajo (Agrawal *et al.*, 1993) para permitir la localización de subsecuencias. Para ello divide las series en regiones, mediante el desplazamiento de una ventana de longitud fija. A cada región se le calculan sus coeficientes DFT. Los puntos característicos de desplazamientos cercanos de la ventana forman una huella debido a que se componen de casi los mismos elementos y se calcula el rectángulo mínimo envolvente o MBR (*Minimum Bounding Rectangle*) que la contiene. Luego estos MBR son almacenados usando árboles $R*$.

Las consultas de rango se basan en el cálculo del punto correspondiente a la secuencia dada, la obtención de los MBR que interseccionan con la región indicada por ese punto y un radio ϵ y, finalmente examinar las subsecuencias a que corresponden los MBR para descartar falsas alarmas.

Continuando esta idea, en (Kahveci and Singh, 2001) el índice se compone de todos los árboles $R*$ con los MBR obtenidos de utilizar diferentes tamaños de ventana deslizante para cada una de las secuencias existentes en la base de datos.

La realización de una consulta de rango, especificada por una secuencia q y un radio ϵ, se realiza dividiendo q en varios trozos haciendo una subconsulta para cada uno de ellos. Los resultados de cada consulta permiten afinar el radio de la siguiente consulta.

Ante el posible problema del aumento de los requerimientos de almacenamiento se propone un mecanismo de compresión del índice que reduce el almacenamiento sin implicar un aumento excesivo del coste de computación.

En (Chan and Wai-chee, 1999) se propone una solución para indexar series basada en la DWT (Discrete Wavelet Transform), concretamente en la transformada de Haar. Las ventajas del uso de ésta se concretan principalmente en ser más fácilmente calculable, proporcionar mayor información acerca de la serie y mejor escalabilidad frente al *índice F*. Además se contempla la similitud de series con desplazamiento vertical.

Como se demuestra en (Wu *et al.*, 2000) la utilización de DWT no reduce el error relativo de coincidencia ni mejora la precisión de las consultas de similitud como se indica en (Chan and Wai-chee, 1999). De esta manera se puede afirmar que la utilización de una transformada u otra proporciona resultados comparables en las búsquedas de similitud en bases de datos de series temporales.

2.2 Reducción de Datos

En los siguientes trabajos se reducen los datos originales de la serie temporal seleccionando directamente un subconjunto de ellos o el resultado de hacer transformaciones en agrupaciones de los mismos.

En (Keogh and Pazzani, 1998) se utiliza una segmentación lineal a trozos de la curva original ya introducida en (Keogh and Smyth, 1997) y basada en el algoritmo presentado en (Pavlidis and Horowitz, 1974). Los segmentos obtenidos se identifican por medio de sus puntos inicial y final, además de un valor que sirve de ponderación relativa del segmento con respecto a la serie. Los segmentos son utilizados en un nueva función de distancia en lugar del conjunto completo de valores de la serie lo que reduce el cálculo. Esta distancia es insensible a traslaciones, tendencias lineales y discontinuidades.

Tanto en (Keogh and Pazzani, 2000) como en (Yi and Faloutsos, 2000) se realiza una reducción directa de la dimensionalidad de la curva con una aproximación constante a trozos, conocida como indexación PCA. La idea es dividir la curva en un una serie de segmentos k abarcando cada uno j puntos consecutivos de la curva original y sustituyéndolos por el valor medio en cada segmento, transformándose en un punto de k-dimensiones.

Se demuestra de forma experimental que la cantidad de series que se manipulan en la fase final de comprobación de falsas alarmas es mucho menor que la proporcionada por el *índice F*.

Mientras que en estos dos trabajos la división de la serie se realiza en trozos idénticos en (Keogh *et al.*, 2001) se utiliza un algoritmo adaptativo utilizando segmentos constantes al que se denomina $APCA$. Al ser los segmentos constantes, la serie original se convierte en una serie de pares de puntos que especifican el valor de los puntos de ese segmento y el extremo derecho de cada segmento. Considerando esta secuencia de pares la descripción de la serie se convierte en un punto en un espacio multidimensional cobre el que se aplican los ya comentados MBR y árboles $R*$.

En el artículo (Keogh and Pazzani, 1999) se aplica el algoritmo de envolvente dinámica de tiempo (Dynamic Time Warping) sobre los datos ya segmentados. Se define así un $SDTW$, o DTW segmentado que, manteniendo todas las propiedades de robustez del DTW, permite una aceleración de las comparaciones muy importante al tiempo que posibilita la búsqueda de subsecuencias.

Por otra parte en (Yi *et al.*, 1998) se aplica el método *FastMap*, introducido en (Faloutsos and Lin, 1995), para generar un índice que permita optimizar las consultas y posteriormente se eliminan las falsas alarmas mediante un comprobación de distancia basada en DTW. De esta manera se obtiene velocidad y una similitud que soporta variaciones locales de la frecuencia de la señal representada por las serie.

Un trabajo que también utiliza el DTW es (Kim *et al.*, 2001). En este caso se define una función de distancia que subestima la distancia DTW y que satisface la desigualdad triangular. Esta función toma como parámetro un vector característico de cuatro elementos obtenido de cada serie. Los elementos del vector para una secuencia son el primer valor, el último, el mayor y el menor.

El mecanismo que se adopta en (Shatkay and Zdonik 1996) para reducir la cantidad de información a almacenar de las series consiste en hacer una división de las series y representar cada subsecuencia por medio de una función. Se sugieren como funciones las curvas Bézier, los polinomios y la interpolación lineal. Con esta forma de sintetización toma una gran relevancia la manera de trocear cada secuencia para permitir posteriormente una comparación congruente.

2.3 Otros Enfoques

Además de las aproximanciones ya comentadas en los siguientes trabajos se utilizan otras propuestas para comparar series temporales que describiremos a continuación.

En (Agrawal *et al.*, 1995a) se utiliza un nuevo concepto de similitud, considerándose que dos curvas son similares si existe un número suficiente de pares de subsecuencias sin solapamiento ordenadas que son similares. Este trabajo crea un índice con pequeñas subsecuencias que representan a todas las secuencias, por medio de su agregación. Así la comparación de dos secuencias se realiza por localizar el conjunto ordenado de subsecuencias atómicas que representa a cada una y comparar cuantas subsecuencias son comunes a ambos conjuntos. Para que dos subsecuencias se consideren similares una debe envolver a la otra con una distancia máxima dada tras ajustar sus escalas.

El trabajo (Keogh and Smyth, 1997) presenta un algoritmo de búsqueda para la localización de patrones en un conjunto de datos. Este algoritmo utiliza como distancia la obtenida de un modelo probabilidades en función de

la información de las características de un serie y de un conocimiento apriorístico de las posiciones relativas de las características individuales. De cada serie se realiza una representación lineal a trozos con la idea de reducir la información que describe a la misma.

Un trabajo que también es interesante comentar es (Cheung and Stephanopoulos, 1990) donde se propone el estudio de series con diferentes escalas desde una perspectiva cualitativa. La idea fundamental es reconocer, partiendo de la serie original, los diferentes elementos característicos de la serie realizando un sucesivo aumento de escala hasta reducir la serie a su máxima simplificación. Con este procedimiento se logra un conocimiento de las tendencias existentes en la serie original.

Finalmente el trabajo de (Jönsson and Badal, 1997) se asemeja al presente trabajo al utilizar el SDL de (Agrawal *et al.*, 1995b) para convertir la serie en una secuencias de símbolos. Posteriormente obtiene una firma de cada serie por medio de una función de dispersión (*hashing*). Esta firma, de menor tamaño que la serie original, es usada para determinar la similitud entre las series.

La principal diferencia de nuestro trabajo respecto a los comentados se centra en que la comparación de las series se hace desde una perspectiva cualitativa de los valores de las mismas, frente al tradicional análisis cuantitativo.

Al mismo tiempo hemos convertido el problema de la comparación de series numéricas a comparación de cadenas con la simplificación que ello representa. Esta modificación en el dominio del problema proporciona la posibilidad de utilización de metodologías muy profusamente estudiadas en los últimos años.

Finalmente la utilización de algoritmos de comparación de cadenas también es una novedad, al tiempo que se mantienen las características de los algoritmos de reconocimiento de voz utilizados en otros trabajos; principalmente la detección de series similares aunque existan desplazamientos en la escala temporal.

3 Lenguaje de Definición de Formas (SDL)

Este lenguaje de definición de formas (Shape Definition Language) propuesto en (Agrawal *et al.*, 1995b) es un muy apropiado para realizar consultas sobre la forma de evolución de valores o magnitudes a lo largo del tiempo.

Dado un conjunto de valores registrados durante un periodo, la primera idea en SDL es dividir el intervalo de los posibles valores de las variaciones entre sucesivos valores de la serie en rangos disjuntos y asignarle a cada uno una etiqueta.

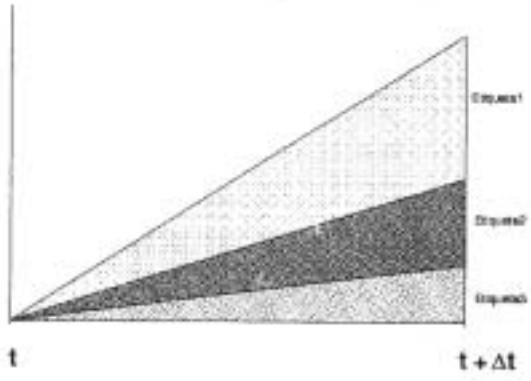

Figure 1: Posible asignación de etiquetas

La figura 1 puede representar una posible división en tres zonas de la parte positiva de los posibles valores de variación y las etiquetas que se le han asignado. El comportamiento de una serie puede describirse considerando las transiciones entre las sucesivas muestras. Se construye una derivada con respecto al tiempo de la serie, calculando la diferencia de amplitud entre las muestras. El valor de esas diferencias se encuadra en uno de los diferentes rangos disjuntos definidos proporcionando una etiqueta del alfabeto.

Así la traducción produce una secuencia de transiciones basada en un alfabeto cuyos símbolos describen la magnitud de los incrementos de los valores de la serie temporal.

Cada símbolo es definido por cuatro descriptores. Los dos primeros son el límite inferior y superior permitidos a la variación entre los valores inicial y final de la transición. Los dos últimos especifican las restricciones sobre los valores inicial y final de la transición.

El alfabeto propuesto en dicho trabajo (Agrawal *et al.*, 1995b) propone la utilización de ocho símbolos diferentes. El tamaño del alfabeto es muy pequeño comparado con los diferentes valores reales que la curva toma a lo largo del tiempo. Una traducción de este tipo permite dar prioridad a la forma sobre los valores originales.

En la figura 2 se muestra un ejemplo de traducción.

Cada cadena de símbolos puede describir un infinito

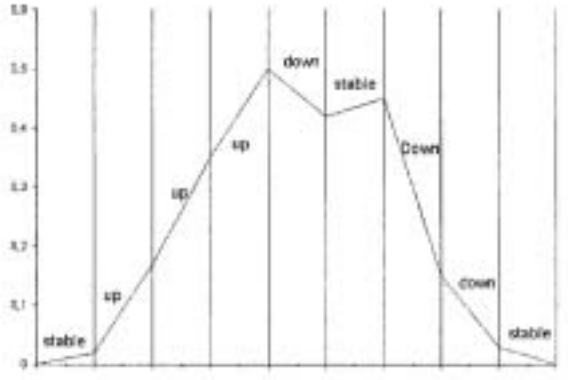

Figure 2: Ejemplo de traducción

número de curvas que cumplan con las restricciones impuestas por los símbolos a las transiciones que representan.

La figura 3 muestra tres curvas diferentes cuya traducción produce la misma secuencia de símbolos; las curvas tienen diferentes puntos iniciales.

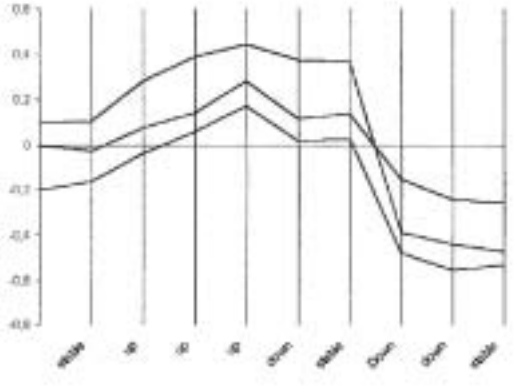

Figure 3: Traducción con idéntica secuencia

El lenguaje SDL considera que cada uno de los símbolos del alfabeto describe una forma elemental y proporciona operadores para poder definir, partiendo de ellas, formas complejas derivadas. Estas formas se pueden generar partiendo de la concatenación, repetición, multiselección y parametrización de formas elementales u otras derivadas.

Finalmente en este trabajo también se propone un tipo de indexación y su mecanismo de almacenamiento para permitir implementación de consultas con los mecanismos de descripción definidos.

Limitamos la utilización de este lenguaje en el presente trabajo a convertir una serie temporal descrita por sus valores numéricos a una cadena de símbolos representando las variaciones entre valores adyacentes.

4 Subsecuencia Común Máxima (LCS)

Trabajando con secuencias de cualquier tipo, desde cadenas de caracteres a secuencias de ADN, una de las medidas de similitud más usadas es la Subsecuencia Común Máxima (LCS) de dos o más secuencias dadas. LCS es la mayor colección de elementos que se encuentran en el mismo orden en dos secuencias distintas .

Sean $S = s_1, s_2, ..., s_m$ y $T = t_1, t_2, ..., t_n$ sobre un alfabeto $\sum$, entonces $LCS(S,T) = U, U = u_1, u_2, ..., u_r$ tal que existen índice $i_1 < i_2 < ... < i_r$, con $1 \leq r \leq m$, y $j_1 < j_2 < ... < j_r$, con $1 \leq r \leq m$, tal que $s_{i_z} = u_z$ y $t_{j_z} = u_z$.

La obtención de la LCS es un problema muy conocido y que pertenece a la clase de problemas NP-Completos, suponiendo que la longitud de las cadenas a tratar es m y n respectivamente, tiene una complejidad $O(m,n)$ en tiempo y espacio utilizando técnicas de programación dinámica.

Desde (Wagner and Fisher, 1974) hasta (Cormode *et al.*, 2001) se han realizado diferentes aproximaciones al problema que intentan reducir la complejidad en alguna medida, ya sea considerando alfabetos limitados, requerimiento lineal del espacio, soluciones aproximadas o incluso algoritmos de acotación por aprendizaje.

Nuestro interés de usar LCS es doble:

- Como el lenguaje SDL produce una cadena de símbolos desde los valores numéricos de la serie temporal, es posible entonces usar este algoritmo para obtener una distancia entre dos series con abstracción en las formas de las curvas.
- El LCS es un caso especial del DWT con la única consideración de la presencia o no de símbolos, con una distancia de 0 ó 1 en cada comparación. Heredando así todas las prestaciones de DWT.

El algoritmo DWT es usado intensamente en el campo del reconocimiento de voz debido a su capacidad de detectar formas similares de ondas que no estén alineadas en el eje temporal. Esta falta de alineamiento induce catastróficos resultados en una comparación con distancia Euclídea.

El fundamento de DWT está en buscar un conjunto de mapeos ordenados entre los valores de dos series, de

forma que se minimice la distancia global o coste de envoltura (warping cost). La idea básica es intentar descubrir que algunos segmentos de una de las series a comparar son muy similares a los de la otra serie sobre los que se han realizado transformaciones de compresión o expansión; es en realidad la búsqueda de variaciones locales de la frecuencia de las series.

5 Índice Cualitativo de Similitud (QSI)

En este articulo se propone incluir conocimiento cualitativo en la comparación de las series temporales. Se propone una medida de similitud basada en la coincidencia de las etiquetas cualitativas que representan la evolución de los valores de las series. Cada etiqueta representa un rango de valores que desde una perspectiva cualitativa podemos entender como similares. Diferentes series con una evolución cualitativamente similar producen la misma secuencia de etiquetas.

La aproximación propuesta obtiene mejores resultados que otros métodos aparecidos en la bibliografía. Por un lado, el uso de toda la información contenida en la serie temporal maximiza la exactitud. Por otro, la consideración de grupos de evoluciones como similares prioriza la focalización de la comparación en la forma general de las curvas y no en sus valores puntuales. En cualquier caso, hay que decir que las series temporales con las que se trabaja se suponen libres de ruido entre muestras donde la evolución se supone lineal y monotónica.

Sea $X = \langle x_0, ..., x_f \rangle$ una serie temporal. La aproximación propuesta en este trabajo se realiza en tres etapas. Primero se realiza una normalización de los valores de la serie temporal, obteniéndose $\tilde{X} = \langle \tilde{x_0}, ..., \tilde{x_f} \rangle$ y a partir de ella se obtiene la serie de diferencias $X_D = \langle d_0, ..., d_{f-1} \rangle$, que se traduce a una cadena de caracteres $S_X = \langle c_1, ..., c_{f-1} \rangle$. La similitud entre dos series temporales se obtiene comparando las dos cadenas obtenidas de la transformación anterior mediante un algoritmo LCS. Esta medida sirve en nuestra aproximación como medida de similaridad entre las series.

Veamos a continuación de manera detallada cada uno de estas transformaciones de manera detallada.

5.1 Normalización

En primer lugar y con la intención de poder comparar cualitativamente las series se realiza una normalización de sus valores al intervalo [0,1].

Sea $X = \langle x_0, ..., x_f \rangle$ una serie temporal. A partir de ella se obtiene la serie temporal normalizada representada como $\tilde{X} = \langle \tilde{x_0}, ..., \tilde{x_f} \rangle$ donde:

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

siendo min y max operaciones que devuelven los valores mínimo y máximo de una secuencia de números.

A partir de esta serie normalizada se obtiene la serie de diferencias $X_D = \langle d_0, ..., d_{f-1} \rangle$ donde

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

Esta serie de diferencias será utilizada posteriormente en el etiquetado para obtener la cadena de caracteres correspondiente a la serie temporal. Es interesante comentar que cualquier $d_i \in X_D$ será un número en el intervalo [-1,1].

5.2 Etiquetado

La normalización propuesta en el apartado anterior considera la evolución de la pendiente en lugar de los valores de la serie. Con la intención de asignar una etiqueta a cada tipo de pendiente, se divide el intervalo de las posibles pendientes en varios grupos y se le asigna a cada uno una etiqueta cualitativa.

La anterior división se realiza de acuerdo con un parámetro δ, proporcionado por los expertos, según el conocimiento que estos tienen sobre el problema. En este sentido, los expertos deben informar sobre que significan las diferentes etiquetas cualitativas en el ámbito del problema, es decir, identificar los rangos en los que se mueven cada una de las etiquetas cualitativas. El valor de este parámetro influye en la calidad de los resultados, sin embargo aún no se ha realizado un estudio detallado sobre su importancia en los resultados ni sobre su relación con las series pertenecientes al dominio

del problema.

Etiqueta	Intervalo	Símbolo
Fuerte incremento	$[1/\delta, +\infty]$	H
Incremento moderado	$[1/\delta^2, 1/\delta]$	M
Ligero incremento	$[0, 1/\delta^2]$	L
Sin variación	0	0
Ligero decremento	$[-1/\delta^2, 0]$	l
Decremento moderado	$[-1/\delta, -1/\delta^2]$	m
Fuerte decremento	$[-\infty, -1/\delta]$	h

En esta tabla, la primera columna representa la etiqueta cualitativa correspondiente a cada intervalo de pendientes, recogido en la segunda columna. La última de las columnas presenta el carácter asignado a cada etiqueta.

El alfabeto propuesto define tres valores para incrementos y otros tres para los decrementos, y uno para la no variación. Es interesante comentar que en este caso, no se hace uso de las restricciones definidas en el SDL.

Este alfabeto es el que se utiliza para obtener la cadena de caracteres asociada a la serie temporal $S_X = \langle c_1, ..., c_{f-1} \rangle$ donde cada c_i representa la evolución de la curva entre dos instantes consecutivos y se obtiene a partir de $X_D = \langle d_0, ..., d_{f-1} \rangle$ asignándole a cada d_i su carácter.

Esta traducción de la serie temporal a una secuencia de símbolos permite abstraernos de los valores de la curva y centrarnos en su forma. Cada secuencia de símbolos describe toda una familia de curvas con una evolución similar.

La figura 4 representa una curva normalizada, los valores de las derivada y las etiquetas asignadas a cada transición entre valores consecutivos. Este ejemplo se ha utilizado un valor $\delta = 5$.

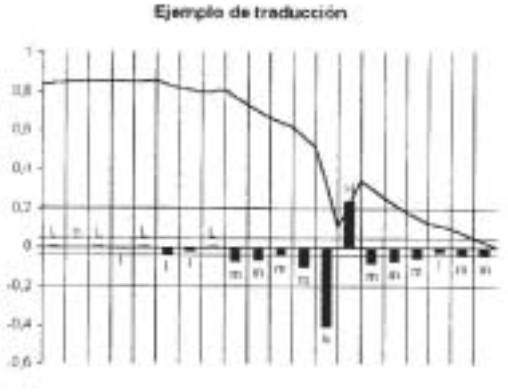

Figure 4: Ejemplo de traducción con QSI

5.3 Definición de Similitud QSI

Sean las series temporales $X = \langle x_0, ..., x_f \rangle$ y $Y = \langle y_0, ..., y_f \rangle$. Normalizando y etiquetando esas series se obtendrán las cadenas S_X, S_Y.

Se define la similitud QSI entre las cadenas como

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

donde ∇S es el cuantificador de conteo aplicado a la cadena S, es decir obtiene la longitud de S. Por otro lado, m se define como $m = max(\nabla S_X, \nabla S_Y)$. De esta manera, se puede interpretar la similitud QSI como el número de símbolos que podemos encontrar en el mismo orden en ambas secuencias dividido por la longitud de la secuencia más larga.

Sean S_X, S_Y, S_Z tres cadenas de caracteres obtenidas de la transformación de tres series temporales. La definición anterior de similitud QSI cumple lo siguiente:
1) El número de símbolos coincidentes y en el mismo orden entre dos cadenas incide directamente en el índice QSI.
2) El tamaño de las cadenas también afecta a QSI. En este sentido, dos cadenas de un tamaño parecido y que comparten un número determinado de símbolos, son más parecidas que otras dos con el mismo número de símbolos pero de tamaños muy diferentes, es decir,

$$\begin{gathered} \nabla S_X \approx \nabla S_Y, \nabla S_X \approx \nabla S_Z, \\ \nabla LCS(S_X, S_Y) \approx \nabla LCS(S_X, S_Z) \\ \Downarrow \\ QSI(S_X, S_Y) > QSI(S_X, S_Z) \end{gathered} \tag{4}$$

3) $S_X \equiv S_Y \Leftrightarrow QSI(S_X, S_Y) = 1$. Esta propiedad indica que la similitud entre cadenas iguales es 1, por la manera en la que se calcula el índice QSI. Y a su vez si el índice es 1, la única posibilidad es que las dos cadenas sean iguales, aunque no tiene porqué serlo las series temporales que originaron ambas cadenas.
4) Se cumple que:

$$\begin{gathered} \text{Si } LCS(S_X, S_Y)) = \emptyset \\ \Downarrow \\ \nabla LCS(S_X, S_Y)) = 0 \\ \Downarrow \\ QSI(S_X, S_Y) = 0. \end{gathered} \tag{5}$$

5.4 Comparativa con Otras Aproximaciones

Para comprobar la mejora de este trabajo respecto a otros aparecidos en la bibliografía, se ha probado la

propuesta presentada siguiendo los mismos pasos que en (Keogh and Pazzani, 1999) realizando el clustering de un conjunto de curvas. Cualquier proceso de clustering agrupa un conjunto de datos en subconjuntos de forma que se maximiza la similitud entre los elementos de un mismo subconjunto y se minimiza la similitud entre diferentes subconjuntos.

En este trabajo se probó en primer lugar sobre el conjunto del lenguaje de signos autraliano del archivo del UCI KDD (Bay, 1999) seleccionando 5 registros para cada palabra. Para hacer posible la comparación de resultados hemos elegido las 10 palabras que se utilizaron en (Keogh and Pazzani, 1999) de entre las 95 palabras incluidas en el archivo. Posteriormente para cada posible emparejamiento de dos palabras distintas (45), se hace un clustering con las 10 secuencias (5 de cada palabra), utilizando un clustering jerárquico por media de grupo, realizando la comprobación utilizando dos medidas. En primer lugar, la distancia definida en el algoritmo DWT clásico aplicado sobre los valores de la serie, obteniéndose que agrupan de forma correcta 22 de los 45 posibles. En segundo lugar, utilizando los índices de similitud QSI propuestos, y aplicándolos a las cadenas de caracteres obtenidas de la traducción de los valores de las series, se agruparon correctamente 44 de los 45 posibles.

El resultado de aciertos obtenido con DWT concuerda con el obtenido en (Keogh and Pazzani, 1999), sin embargo el que se obtiene con índice de similitud QSI presentado en este trabajo es muy superior.

6 Modelo de Crecimiento Logístico con Retraso

Es común encontrar en el mundo real procesos de crecimiento donde una fase inicial de crecimiento exponencial es seguida por otra fase de acercamiento asintótico a un valor de saturación (figura 5.a). A estos procesos se le dan los nombres genéricos: logístico, sigmoidal, o *s-shaped*. En la literatura, estos modelos se han estudiado profusamente. Abundan en procesos naturales, sociales y socio-tecnológicos. Por citar algunos aparecen en la evolución de las bacterias, en la extracción de mineral, en el crecimiento de la población mundial, en desarrollos económicos, también las curvas de aprendizaje, determinados fenómenos dentro de una población, como rumores o epidemias o un nuevo producto que se introduce en el mercado también muestran este tipo de comportamiento. Estos comportamientos se muestran en la figura 5.b. El patrón de comportamiento es bimodal hacia dos atractores: A crecimiento normal, y O decadencia y extinción.

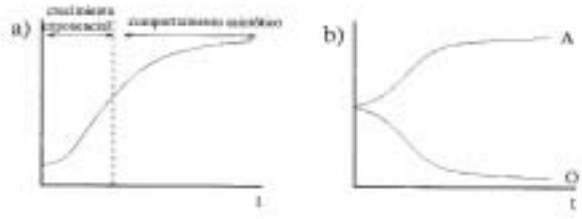

Figure 5: Modelo con crecimiento logístico

Cuando se añade un retraso en el camino de realimentación, entonces las ecuaciones diferenciales del modelo S son

$$\Phi \equiv \begin{cases} \dot{x} = x(n\,r - m), \\ y = delay_r(x), \\ r = h(y), \\ x > 0 \end{cases} \tag{6}$$

siendo n el factor de crecimiento, m el factor de decrecimiento, y h una función con un máximo en la parte positiva de x e y. Las condiciones iniciales son

$$\Phi_0 \equiv \begin{cases} X_0 \in [LP_x, MP_x], \\ LP_x(m), LP_x(n), \\ r \in \{MP_r, VP_r\} \end{cases} \tag{7}$$

donde LP, MP, VP son operadores unarios cualitativos para las variables x, r, definidos de acuerdo con (Ortega *et al.*, 1999).

La metodología (Ortega *et al.*, 1999) se aplica para obtener la base de datos con las series temporales resultantes de la evolución del modelo. En esta metodología lo que se hace es transformar el conjunto de ecuaciones diferenciales del modelo (6) en la siguiente familia de modelos cuantitativos.

$$\begin{cases} \dot{x} = x(n\,r - m), \\ y = delay_r(x),\ x > 0,\ r = H_1(y) \\ H_1,\ x_0 \in [0, 3], \\ m, n \in [0, 1],\ r \in [0.5, 10] \end{cases} \tag{8}$$

La descripción detallada de cómo se ha obtenido (8) se encuentra en el citado artículo.

La selección y simulación de modelos cuantitativos de esta familia permite obtener la base de datos con las series temporales resultantes de la evolución del modelo. Sobre esta base de datos queremos obtener los diferentes comportamientos que tiene aplicando para ello el índice de similitud QSI. La matriz resultante de aplicar este índice se recoge en la figura 6.

Matriz de Similitud - 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	0,366	0,384	0,35	54X
		0,294	0,316	0,39	0,384	0,35	55X
			0,756	0,792	0,566	0,586	1X
				0,754	0,58	0,562	18X
					0,632	0,62	77X
						0,93	17X

Figure 6: Matriz QSI del modelo logístico

En esta matriz se observan tres posibles comportamientos, de acuerdo con el QSI obtenido. Los resultados

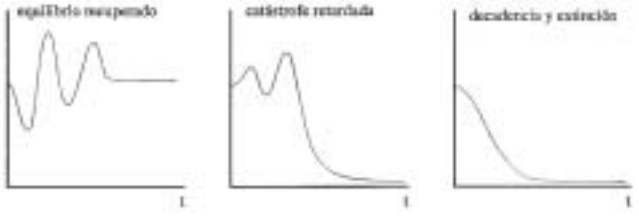

Figure 7: Modelo de crecimiento logístico con retraso

obtenidos están en concordancia con aquellos que se obtienen cuando un razonamiento matemático se lleva fuera (Aracil *et al.*, 1997).

En las figuras 8, 9 y 10 se representan gráficamente las series temporales agrupadas según estos comportamientos. La figura 8 corresponde a las gráficas de las series 1x,18x y 77x cuyo comportamiento se clasifica como *equilibrio recuperado*. De igual manera, la figura 9 representa las series 17x y 73x correspondientes al comportamiento etiquetado como *catástrofe retardada* y finalmente la figura 10 correspondiente al comportamiento etiquetado como *decadencia* que corresponde a las series temporales 50x, 54x y 55x.

7 Conclusiones y Trabajo Futuro

En este articulo se ha presentado un índice para medir la similitud de series temporales atendiendo a sus características cualitativas. Además la mejora propuesta ofrece unos resultados mejores comparándola con algoritmos existentes en la bibliografía y cuyos costes computacionales son comparables.

El índice de similitud cualitativa QSI presentado utiliza en primer término una normalización de la serie

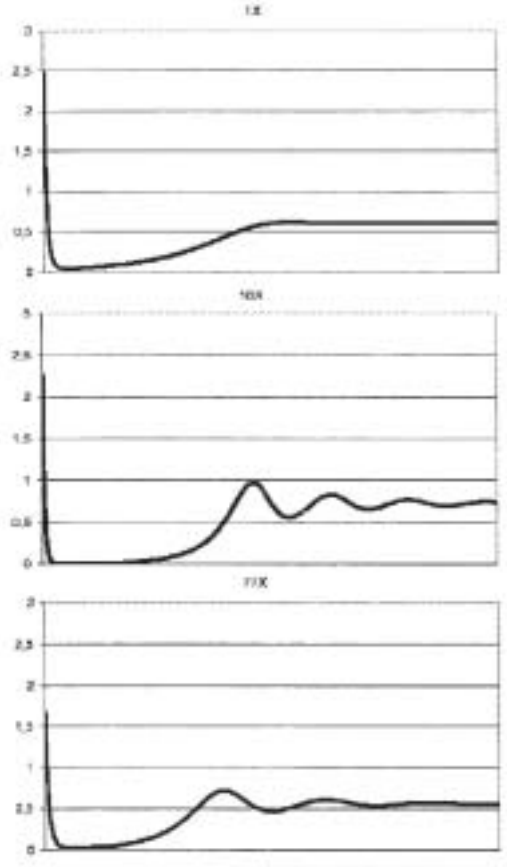

Figure 8: Equilibrio recuperado

temporal, para a partir de la serie normalizada, obtener otra como resultado del cálculo de la diferencia. Finalmente partiendo de esta última y de un alfabeto definido cualitativamente obtener una cadena de caracteres. Se aplican algoritmos de LCS para calcular este índice de similitud.

Los resultados obtenidos concuerdan con otros aparecidos en la bibliografía, si bien mejoran las clasificaciones que éstos realizan.

En cuanto al trabajo futuro cabe decir que, en primer lugar pretendemos optimizar el mecanismo de división de la pendiente en zonas, estudiando la elección del número de regiones y sus puntos de corte. Estamos interesados en encontrar un algoritmo que pueda proporcionar para cada dominio de aplicación, y con un aprendizaje inicial con un subconjunto de series, los valores más idóneos en lo referente al número de rangos y el intervalo de cada uno. Es posible que los resultados de ese trabajo permitan realizar optimizaciones en la obtención del índice presentado.

Otras líneas de investigación futuras son la aplicación

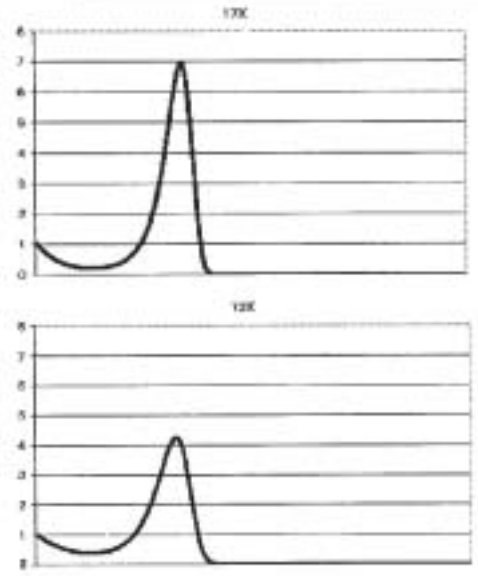

Figure 9: Catástrofe retardada

de esta técnica a series temporales con ruido y la posibilidad de definir grados de similitud, atendiendo a modelos con diferentes escalas de tiempo. Además es necesario hacer un estudio sobre la posibilidad de modificar la mejora propuesta para que se añada la información de la distribución de los segmentos similares de las dos secuencias con la intención de proporcionar un mayor grado de exactitud en la obtención del grado de similitud.

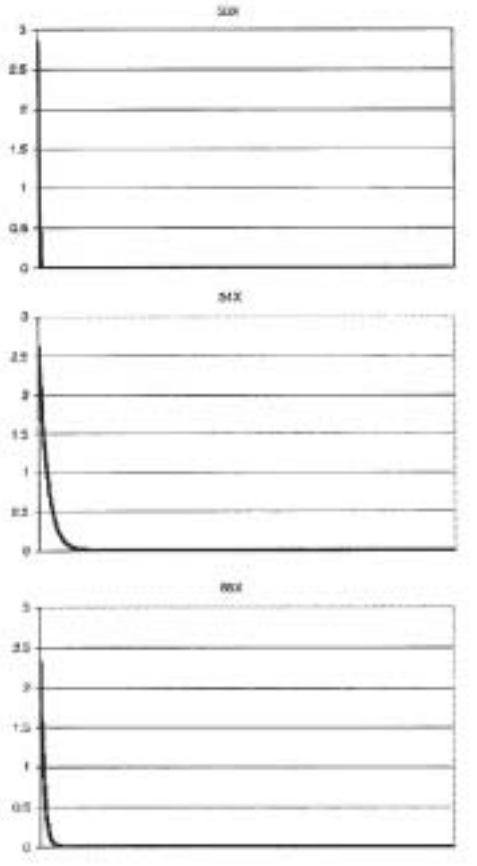

Figure 10: Decadencia y extinción

Referencias

Agrawal R., Faloutsos C. y Swami A., Efficient similarity search in sequence databases. In *Proc. of the Fourth Intl. Conf. on Foundations of Data Organization and Algorithms (FODO '93)* , Chicago ,1993.

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

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

Apostolico A., String Editing and Longest Common Subsequences. *Handbook of Formal Languages* vol. 2 Linear Modeling: Background and Application, 1997, pp.361-398.

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

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

Beckmann N., Kriegel H.-P., Schneider R. and **Seeger B.**, "The r*-tree: an eficient and robust access method for points and rectangles", ACM SIGMOD, 1990, pp. 322-331.

Chan K. and **Wai-chee F.A.**, Efficient time series matching by wavelets *Proc.* 15th *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), 1990, pp. 511-539.

Cormode G., Muthukrishnan S., Paterson M., Sahinalp S.C. and **Vishkin U.** , Techniques and applications for approximating strong distances - Rough Draft, *http://citeseer.nj.nec.com/320221.html* , 2001.

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

Faloutsos C. and **Lin K.I.**, Fastmap: A fast algorithm for indexing, data-minig and visualization of traditional and multimedia datasets. *The ACM SIGMOD Conference on Management of Data*, San Jose, 1995.

Goldin D.Q. and **Kanellakis P.C.**, On similarity queries for time-series data: constraint specification and implementation. In *1st Intl. Conf. on the Principles and Practive of Constraint Prog.*, Minneapolis, 1994, pages 419-429.

Jönsson H.A. and **Badal D.Z.**, Retrieval of one-dimensional data *First European Symposium on Principles of Data Mining and Knowledge Discovery*, 1997.

Kahveci T. and **Singh A.** , Variable length queries for time series data. In *Proceedings of the* 17^{th} *Intl. Conf. on Data Engineering*, Heidelberg, 2001.

Keogh E.J. and **Smyth P.**, A probabilistic approach to fast pattern matching in time series databases, *Proceedings of the* 9^{th} *International Conference on Tools with Artifial Inteligence*, IEEE Press, 1998, pp. 578-584.

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* , AAAI Press, 1998,pp. 239-241,.

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, In *Proc. of the Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 2000.

Keogh E.J., Chakrabarti K., Mehrotra S. and **Pazzani M.J.**, Locally adaptative dimensionality reduction for indexing large time series databases, ,*SIGMOD*, 2001.

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

Oppenheim V. and **Schafer R.W.**, *Digital Signal Processing* , Prentice Hall, Englewood Cliffs, N.J., 1975.

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

Ortega J.A., Gasca R.M., and **Toro M.** , Searching for similar semiqualitative temporal patterns in time-series databases. In the 14^{th} International Workshop on Qualitative Reasoning. Morelia(Mexico), 2000, pp. 111–122.

Pavlidis T. and **Horowitz S.**, Segmentation of plane curves. *IEEE Transactions on Computers*, Vol.C-23,No.8,1974.

Raifei D. and **Mendelzon A.**, Similarity-based queries for time data series. In *Proc. of the ACM SIGMOD Intl. Conf. of Management of Data(SIGMOD '97)*, Tucson, 1998, pp. 13-24.

Raifei D. and **Mendelzon A.**, Efficient Retreival of similar time sequences using DFT. In *Proc. of the* 5^{th} *Intl. Conf. on Foundations of Data Organization and Algorithms (FODO '98)*. Kobe, 1998.

Raifei D., On Similarity-based queries for time series data. In *Proc. of the* 15^{th} *International Conference on Data Engineering.* , Sydney, 1999.

Shatkay H. and **Zdonic S.**, Approximate queries and representation for large data sequences. In *Proc. of the* 12^{th} *International Conference on Data Engineering.* , 1996, pp. 546-553.

Wagner R.A. and **Fisher M.J.**, The string-to-string correction problem. *Journal of the ACM 21*, 1974, pp. 168-173.

Wu D.,Agrawal D. and **Abbadi A.**, A comparison of DFT and DWT based Similarity Search in Time-Series Databases. *Proc. of the* 9^{th} *International Conference on Information and Knowledge Management*, 2000.

Yi B.K., Jagadish H. and **Faloutsos C.**, Efficient retrieval of similar time sequences under time warping. *IEEE Intl. Conf. on Data Engineering*, 1998,pp. 201-208.

Yi B.K. and **Faloutsos C.**, Fast time sequence indexing for arbitrary L_p norms. *Proceedings of the* 26^{th} *Intl. Conf. on Very Large Databases*,Cairo, 2000.

Juan Antonio Ortega *nacido en 1968 es Ingeniero en Informática desde 1992 y Doctor en Informática desde el año 2000, ambos títulos obtenidos en la Universidad de Sevilla (España). Es profesor del Departamento de Lenguajes y Sistemas Informáticos de la citada Universidad desde el año 1992. Su principal campo de investigación se centra en la simulación de sistemas dinámicos con conocimiento semicualitativo y la obtención de sus patrones de comportamiento temporal.*

Francisco Javier Cuberos *obtuvo su título de Ingeniero en Informática en 1993 en la Universidad de Sevilla (España). Trabaja como Administrador de Sistemas en RadioTelevision de Andalucía (R.T.V.A.) desde 1991. Su investigación se centra en el análisis de series temporales en sistemas dinámicos.*

Rafael M. Gasca *obtuvo el título de Doctor en Informática en 1998 en la Universidad de Sevilla en España. Es profesor del Departamento de Lenguajes y Sistemas Informáticos de la Universidad de Sevilla desde 1991. Sus principales áreas de investigación son la programación con restricciones y el razonamiento semicualitativo.*

Miguel Toro *obtuvo el título de Doctor en Ingeniería en 1987 en la Universidad de Sevilla en España. Es profesor desde 1985 y Director del Departamento de Lenguajes y Sistemas Informáticos desde 1993 y desde el año 2000 es Director de la Oficina de Transferencia de Resultados de la Investigación de la Universidad de Sevilla. Sus líneas de investigación se centran en la Ingeniería del Software, la simulación de sistemas dinámicos y los métodos formales.*