Similitud de series de tiempo basada en longitud de patrones de la transformada por aproximación móvil

Francisco Javier Garcia-Lopez, Ildar Batyrshin, Alexander Gelbukh

Instituto Politécnico Nacional, Centro de Investigación en Computación, México

b151153@cic.ipn.mx, batyr1@gmail.com, gelbukh@gelbukh.com

Resumen. Se propone medir la similitud de series de tiempo utilizando la longitud de patrones que arroja la Transformada por Aproximación Móvil. Además, se propone un método para la selección automática de ventanas de tiempo y un método para la visualización de los intervalos de tiempo en los cuales las similitudes entre dos series ocurren. El método se aplica a series de tiempo de acciones de empresas obtenidas del sitio de *Google Finance*.

Palabras clave: Series de tiempo, medida de similitud, MAT, selección automática de ventana de tiempo.

Similarity of Time Series based on the Length of the Patterns of the Moving Approximation Transform

Abstract. We propose to measure the similarity of time series using the length of the patterns given by the Moving Approximation Transform. In addition, we propose a method for automatic selection of the time window for this transform and a method for visualization of the time intervals in which the similarities between the two series occur. The method is applied to time series of the stock values of the firms obtained from the Google Finance site.

Keywords: Time series, similarity measures, MAT, automatic selection of time window.

1. Introducción

Una de las tareas que ha tenido más atención en los últimos años en el análisis de bases de datos de series de tiempo es la medición de similitud entre series de tiempo [1]. Varios métodos se han desarrollado en minería de datos de series de tiempo para medir tal similitud [3–5]. En la siguiente sección se abordan algunos conceptos que se han desarrollado para esto.

La Transformada por Aproximación Móvil (MAT por sus siglas en inglés) remplaza los valores de una serie de tiempo por los valores de pendiente de las líneas que aproximan sub-secuencias de la serie [2]. La aproximación se hace por regresión lineal y se utilizan ventanas de tiempo para indicar el número de puntos que se aproximan. La asociación de tendencias locales y distancia de tendencias locales son medidas que se obtienen utilizando la MAT. Estas medidas tienen la propiedad de ser invariantes bajo normalizaciones o transformaciones lineales.

La selección de la ventana de tiempo para la transformada MAT es importante para producir los resultados más significativos. En el presente artículo se seleccionan las ventanas de tiempo buscando que las que brinden mayor información.

Batyrshin et al. [2, 7] formularon el problema de desarrollar medidas de asociaciones negativas entre series de tiempo, esto es, cuando una serie tiene una tendencia a la alza en el mismo periodo que otra serie tiene tendencia a la baja. Por ejemplo, dos empresas competidoras pueden tener dinámicas inversas en la bolsa si el precio de las acciones de una sube al mismo tiempo que el precio de las acciones de la otra baja. En [2] fueron propuestos medidas de asociación de tendencias locales para encontrar asociaciones positivas y negativas entre series de tiempo; además se incluyeron diversos ejemplos de su uso para análisis de asociaciones entre datos financieros, económicos, políticos, etcétera.

Específicamente, en [7] se propuso un método de análisis de asociaciones entre patrones de series de tiempo que tienen asociaciones positivas o negativas basándose en la técnica de tendencias locales. El problema de la búsqueda de estos patrones aparece cuando las asociaciones positivas y negativas entre series de tiempo cambian en el tiempo, por ejemplo cuando los precios de la acción pueden cambiar dependiendo de eventos económicos como el lanzamiento de un nuevo producto, o la alianza de varias empresas, baja de precio de petróleo, etc.

En el presente artículo se propone un método de visualización de patrones de las asociaciones, una medida de similitud entre series de tiempo con diferentes tamaños de patrones. Además se presentan las series en forma de red asociativa y finalmente una forma de selección de ventana de tiempo. En [2] se utilizó la medida coseno para medir similitud después de aplicar la transformación MAT pero en el presente artículo se mide similitud basándose en los patrones continuos en que dos series tienen asociaciones negativas o positivas, utilizando el total de patrones y la suma de sus asociaciones, además se propone un método de visualización de los periodos en que se mantienen las asociaciones.

El resto del artículo se estructura de la siguiente manera. En la sección 2 se hace una revisión de los trabajos relacionados. En la sección 3 se presenta la medida que se propone basada en longitud de patrones. En la sección 4 se describe la metodología para seleccionar automáticamente la ventana de tiempo. En la sección 5 se muestran los resultados obtenidos. Finalmente, la sección 6 concluye el artículo.

2. Trabajos relacionados

En esta sección discutimos los trabajos que abordan el problema de la medición de la similitud entre los series de tiempo. Por las restricciones del espacio, omitimos la

revisión del estado del arte en la selección de las ventanas de tiempo y la visualización de las asociaciones.

Algunos trabajos utilizando distancia euclidiana miden similitud comparando sub-secuencias de las series sin importar que éstas no estén alineadas en tiempo, esto es, que el i-ésimo punto de una serie corresponde con el i-ésimo punto de la otra. Para esto se elige un parámetro que indica qué tanto desplazamiento está permitido, entre mayor sea el parámetro la búsqueda se hace más lenta.

Das et al. [3] proponen un algoritmo aleatorizado basado en programación dinámica para calcular similitud utilizando la sub-secuencia común más larga (LCSS por sus siglas en inglés). El algoritmo toma en cuenta desplazamiento en tiempo, cuya máxima tolerancia está dada por un número entero positivo δ. Además de una transformación lineal que permite comparar series con diferentes valores base (por ejemplo, una serie que varía alrededor del valor 100 y otra que varía alrededor del valor 30) y diferentes escalas. El umbral de tolerancia está dado por un número real ε que toma valores entre cero y uno. Tomando las anteriores consideraciones en cuenta, la medida de similitud se da por l/n, donde l es la longitud de la sub-secuencia común más larga y n es la longitud total de la serie.

Alcock et al. [4] miden similitud basándose en características. Las características que se extraen son clasificadas como de primer y de segundo orden. Las características de primer orden son: media, desviación estándar, asimetría (*skewness* en inglés) y la curtosis. Las características de segundo orden son energía, entropía, correlación inercia y homogeneidad local. Estas últimas características fueron consideradas por su uso en imágenes, por lo que las series de tiempo se transformaron en matrices bidimensionales, para esta transformación primero se hace una cuantización Q de los valores, por ejemplo, si se hace una cuantización de la serie 1, 2, 3, 4 con dos niveles, la serie quedaría de la siguiente forma: 1, 1, 2, 2. El segundo paso es la construcción de la matriz $\mathrm{c}(\mathrm{i},\mathrm{j})$ donde el punto (i,j) representa el número de veces que un número en la serie con nivel i es seguido por un punto con nivel j a una distancia d_1. Los parámetros que mostraron los mejores resultados de acuerdo a los experimentos de los autores fueron $\mathrm{Q} = 3$ y $\mathrm{d}_1 = 1$. Además de las antes mencionadas se usan otras características de segundo orden. Para ello se genera el arreglo unidimensional donde cada valor del arreglo en la posición i es la diferencia entre el valor en i y el valor en $i + d_2$. Este nuevo arreglo tendrá una longitud de máximo $n - d_2$. Donde n es la longitud de la serie y d_2 es un parámetro a variar. De este nuevo arreglo se obtienen las mismas medidas que las de la serie de primer orden.

Lin et al. [5] consideran la distorsión dinámica temporal (DTW por sus siglas en inglés). La DTW además del desplazamiento en tiempo, como el considerado en LCSS, también considera la velocidad, es decir, si una serie cambia más rápido que la otra pero de la misma forma. Para implementar DTW se construye una matriz con las distancias de cada punto de una serie contra cada punto de la serie con la que se compara y se busca, utilizando programación dinámica, el camino en la matriz que minimice su distancia acumulativa, esto es, la distancia óptima que minimice la distorsión. En el cálculo del camino óptimo se suelen establecer restricciones alrededor del camino original que establecen qué tan lejos se permite hacer la búsqueda del camino óptimo para acelerar el cálculo. Las restricciones más comunes son: la banda de Sakoe-Chiba y el paralelogramo de Itakura.

Las referencias de [3–5] se enfocan en medir distancia, sin embargo, hay otra tarea importante en series de tiempo es la reducción de dimensiones. En ocasiones las series de tiempo son tan grandes que aplicando directamente los algoritmos conocidos sería muy costoso computacionalmente. Los métodos más conocidos para reducir dimensiones son: transformada discreta de Fourier (DFT por sus siglas en inglés), descomposición en valores singulares (SVD por sus siglas en inglés) o transformada discreta de ondícula (DWT por sus siglas en inglés: *discrete wavelet transformation*).

Finalmente, Ye at al. [6] utilizan la distancia euclidiana y el coeficiente de correlación de Spearman para comparar mediciones de sensores de vibración de dos diferentes fabricantes. Se combinan ambas mediciones de tal forma que se tienen cuatro clases: las series que tienen coeficiente Spearman alto y distancia pequeña son similares, coeficiente Spearman bajo y distancia grande son disimilares, distancia grande y coeficiente Spearman alto significa que los rangos son parecidos pero diferente escala, por ejemplo, series similares pero con algunos puntos con valores muy alejados entre ellas; distancia pequeña y coeficiente Spearman bajo se considera que tiene ruido que varía en un rango reducido afectando el rango de los puntos de la serie pero no la distancia. Los datos que arrojan los sensores tienen la misma longitud y frecuencia de muestreo.

3. Marco teórico

Una serie de tiempo [7] de longitud n, donde n es un entero positivo, es una secuencia de números reales $x = (x_1, x_2, \dots, x_n)$ correspondientes a puntos en el tiempo $t = (1, 2, \dots, n)$. Una serie de tiempo puede ser denotada simplemente como x. Una ventana de tiempo W_i de longitud $k > 1$ es una secuencia de índices $W = (i, i+1, \dots, i+k-1), i \in \{1, \dots, n-k+1\}$. Se define $x_{W_i} = (x_i, x_{i+1}, \dots, x_{i+k-1})$ a la secuencia de valores de ventana de tiempo correspondientes a la serie x. Una secuencia $J = (W_1, W_2, \dots, W_{n-k+1})$ de todas las ventanas posibles de tamaño k para $1 < k \leq n$ es llamada ventana deslizante.

Una función $f_i = a_i t + b_i$ con parámetros $\{a_i, b_i\}$ que minimice la ecuación

$$Q(f_i, x_{W_i}) = \sum_{j=i}^{i+k-1} (f_i(t_j) - x_j)^2 = \sum_{j=i}^{i+k-1} (a_i t_j + b_i - x_j)^2 \qquad (1)$$

es una aproximación de mínimos cuadrados de x_{W_i}, o regresión lineal. Los valores a_i, b_ise calculan de la siguiente manera:

$$a_i = \frac{\sum_{j=i}^{i+k-1} (t_j - \overline{t_\iota})(x_j - \overline{x_\iota})}{\sum_{j=i}^{i+k-1} (t_j - \overline{t_\iota})^2}, \quad b_i = \overline{x_\iota} - a_i \overline{t_\iota}, \qquad (2)$$

donde:

$$\overline{t_i} = \left(\frac{1}{k}\right) \sum_{j=i}^{i+k-1} t_j \quad \text{y} \quad \overline{x_i} = \left(\frac{1}{k}\right) \sum_{j=i}^{i+k-1} x_j \,. \tag{3}$$

La transformación queda de la forma: $a = a_1, a_2, \ldots, a_m$, donde $m = n - k + 1$. Una propiedad importante de la transformada es que es invariante a transformaciones lineales aplicadas simultáneamente. Esto es,

$$MAT(rx + s, ry + s) = MAT(x, y), \quad r, s \in \mathbb{R}. \tag{4}$$

Se define la asociación entre dos transformadas $a_1 = (a_{1_1}, a_{1_2}, \ldots, a_{1_m})$ y $a_2 = (a_{2_1}, a_{2_2}, \ldots, a_{2_m})$ al producto término por término. En el presente trabajo se consideró únicamente el signo de las pendientes por lo que la asociación será definida de la siguiente manera:

$$A(a_1, a_2) = \left(sgn\left(a_{1_1} \cdot a_{2_1}\right), sgn\left(a_{1_2} \cdot a_{2_2}\right), \ldots, sgn\left(a_{1_m} \cdot a_{2_m}\right)\right). \tag{5}$$

Un patrón positivo (negativo) es una sub-secuencia continua de signos positivos (negativos) con la longitud más larga [7].

4 Selección de la medida

En este artículo, proponemos un método para medir similitud basada en longitud de patrones. La primera aproximación para llegar a esta medida fue considerar un porcentaje de la asociación de patrones que se va mostrar, por ejemplo, el 15% de las asociaciones más grades. El problema de este método es que se excluyen patrones que pueden estar muy cerca del umbral y que podrían ser considerables para medir la similitud entre las series. Por lo anterior se busca medir robustamente la similitud de patrones y seleccionar el tamaño de ventana para que tenga mayor confianza.

Para medir similitud se utiliza la longitud de los patrones entre dos series de tiempo en lugar de la medida coseno que se utilizó en [2] para medir su similitud positiva o negativa. Se considera tanto el número total de patrones como la suma de estos. El principal inconveniente de utilizar la medida coseno es que si el número de pendientes positivas es el mismo que de pendientes negativas es resultado es cero (al considerar las el signo de las pendientes solamente).

Para obtener la lista de patrones entre dos series de tiempo se obtiene el signo de la pendiente, de cada serie y se obtiene la multiplicación entre ellas. Obteniendo así un patrón positivo cuando ambas pendientes son positivas o ambas negativas y un patrón negativo cuando tienen signo diferente. Es decir, positivo cuando hay una asociación positiva y negativo en cuando hay asociación negativa. En la fig. 1 se muestran asociaciones para una ventana de 30 entre dos empresas petroleras.

La fig. 2 muestra la forma propuesta de visualizar las asociaciones utilizando sólo el signo de la pendiente. Los intervalos continuos de valor +1 (–1) forman un patrón positivo (negativo). Por ejemplo, en la fig. 2 tenemos la secuencia de patrones de longitud (10, -3, 57, -2, 10, -2, 97, -3, 15, -2, 22), donde los patrones negativos llevan el signo menos.

El gráfico llega hasta noviembre pues se grafican los puntos iniciales de las pendientes. La medida de similitud por longitud de patrones se aplica a la lista de patrones positivos y negativos. Una vez que se tiene la lista de patrones se obtiene su suma y su longitud y se les considera de acuerdo a la siguiente fórmula:

$$\mathrm{SIM}(x) = \frac{\mathrm{sum}(x)}{k_{max} - k + 1} \cdot \left(\frac{\mathrm{ceil}\left(\frac{k_{max} - k + 1}{2}\right) - \mathrm{len}(x)}{\mathrm{ceil}\left(\frac{k_{max} - k + 1}{2}\right) - 1} \right). \quad (6)$$

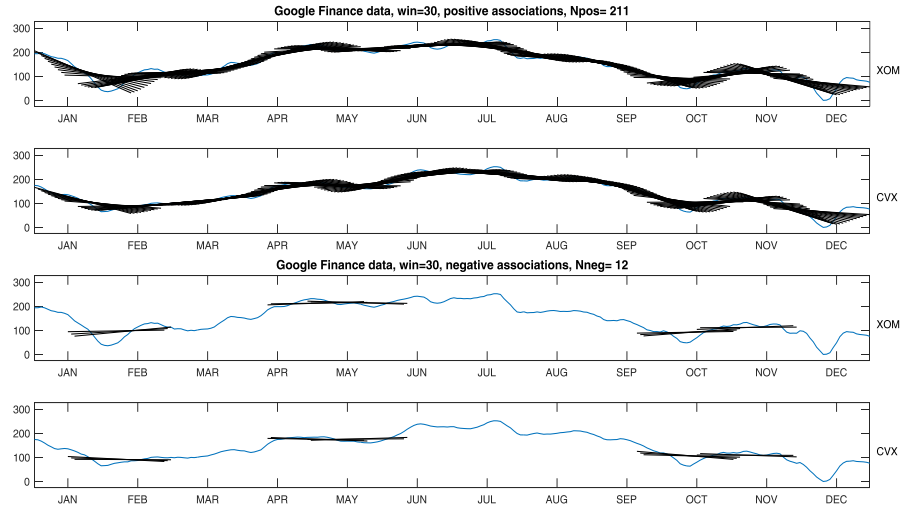

Fig. 1. Asociación entre Chevron y ExxonMobil con ventana de tiempo *k* = 30, datos de 2014.

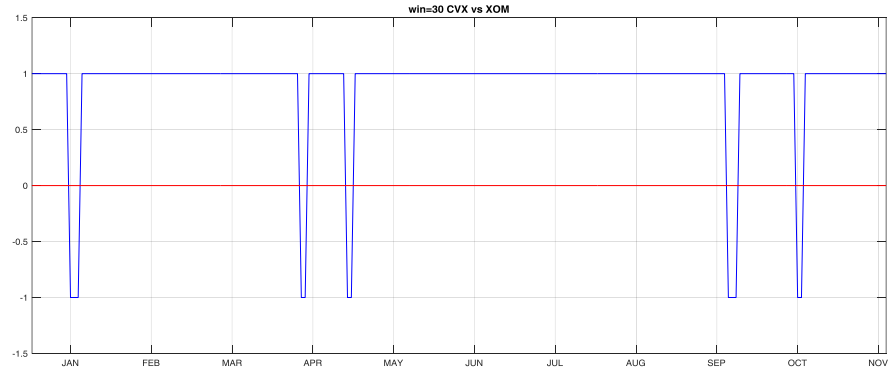

Fig. 2. Los patrones positivos para esta asociación son: {10, 57, 10, 97, 15, 22} mientras que los patrones negativos son: {3, 2, 2, 3, 2}.

En (6) x es la lista de patrones, k es el tamaño de la ventana, k_{max} es el máximo valor que puede tomar la ventana, sum(x) es la suma de todos los valores del patrón,

len(x) es el tamaño de la lista o longitud del patrón y la función ceil asigna el entero positivo más pequeño mayor al número dado.

El primer factor es la suma de los patrones normalizada con respecto a la suma máxima. El segundo factor es la longitud del patrón invertida (afecta negativamente a la similitud) y normalizada. Entre mayor sea la longitud el patrón estará más fragmentado y las series son menos similares ej. El patrón {5} indica mayor asociación que el patrón {1, 1, 1, 1, 1} pues aunque la suma de ambos es la misma en el primero asociaciones están seguidas. Se nota que mientras el tamaño de la ventana se hace más grande, el denominador del primer factor se hace más pequeño esto es porque entre más grande la ventana menor es el número de pendientes que arroja la transformación MAT.

5 Selección de la ventana

Una ventana de tamaño k indica que cada regresión lineal para el cálculo de la MAT será de k puntos, permitiendo un total de $n - k + 1$ ventanas sobre la serie. Se mostró en [2, 7] que ventanas pequeñas detectan los cambios más sensibles, mientras que ventanas más grandes detectan cambios en intervalos de tiempo mayores. Pero la selección de la ventana no se definió con certeza, la cual es una tarea de importancia ya que como se verá, buscando las ventanas apropiadas se pueden obtener una mejor interpretación de las asociaciones.

Como ya se mencionó en la introducción la selección de la ventana de tiempo es importante para saber qué información se está obteniendo. En este trabajo hace la búsqueda de ventanas cuyo valor se ubique entre 2 y la cuarta parte de la mayor ventana posible, lo cual fue determinado experimentalmente, pues con ventanas mayores las series se van haciendo más similares por el hecho de que el número de pendientes que se comparan va disminuyendo. Lo anterior se ejemplifica en la fig. 3.

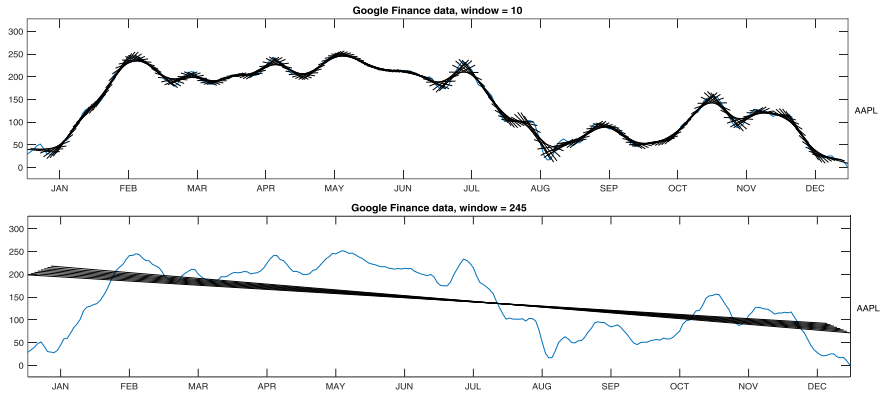

Fig. 3. Comparación del número de pendientes para ventana grande y ventana pequeña.

Se puede afirmar que si dos series están asociadas en una ventana de tiempo pequeña, su similitud es más significativa que cuando el mismo valor de asociación se presenta en ventanas grandes. Las ventanas de tiempo entre más grandes miden sólo como termina una serie con respecto a cómo comienza, es decir si incrementa o disminuye.

El siguiente paso es la selección de la ventana de tiempo. Se propone seleccionarla a partir los valores máximos de los patrones y la longitud de éstos.

Del diagrama como el mostrado en la fig. 4 se obtiene una lista de qué par de empresas es la que tiene el valor más alto para cada valor de ventana (valor máximo por ventana). También se obtiene por cuántas ventanas consecutivas se mantienen ese par como el más alto (longitud del par). Los valores máximos dan información de entre todos los pares, cuáles son los más significativos dado cierto tamaño de ventana, sólo se toman en cuenta los valores mayores a 0.5.

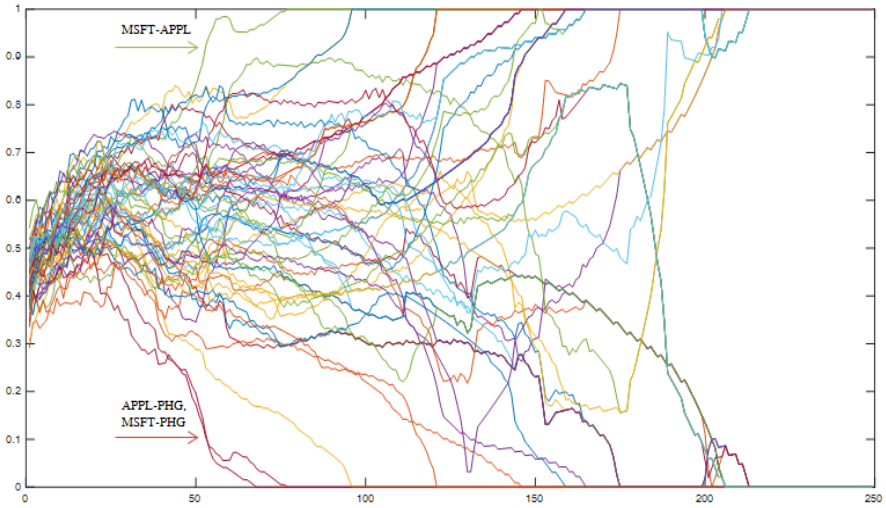

Fig. 4. Similitud positiva utilizando diversos valores de ventana. Empresas de Tecnologías de la información en 2014.

Para sugerir una ventana se le da un peso mayor a la longitud del par que al valor máximo. Una mayor longitud refleja que una similitud dominante de un par de series de tiempo se mantiene para más tamaños de ventana consecutivos. Otra razón para elegir la longitud del patrón es cubrir un mayor intervalo de tiempo. Cabe mencionar que el inconveniente de esta elección es que puede haber otros patrones por debajo del máximo que podrían sugerir otras ventanas además del máximo. Para generar el arreglo de ventanas sugeridas se multiplican los valores de longitud y de valor máximo y el valor mayor de éstos se toma como primera sugerencia para ventana y sucesivamente se hacen las demás sugerencias.

En la fig. 4 se grafica como cambian los valores de similitud de patrones con respecto al tamaño de ventana. Se observa que conforme el valor de la ventana aumenta los patrones se van asociado y quedan ya sea en +1 o 0. Esto se debe a que, como ya se mencionó, los valores mayores de ventana reflejan sólo como empieza y termina una serie contra como empieza y termina la serie con la que se está comparando. Cerca del punto 50 se observa cómo dos series se acercan a cero mientras que una se acerca a 1, esta información es en parte redundante (refleja el hecho de que si dos series son similares entonces si una de ellas es disimilar a una tercera, la otra también lo será) ya que la serie que se acerca a 1 es la de MSFT-APPL, mientras que las que se acercan a

cero son APPL-PHG y MSFT-PHG. Este tamaño de ventana refleja sólo como unas series crecen y otras decrecen gradualmente como se observa en la fig. 5.

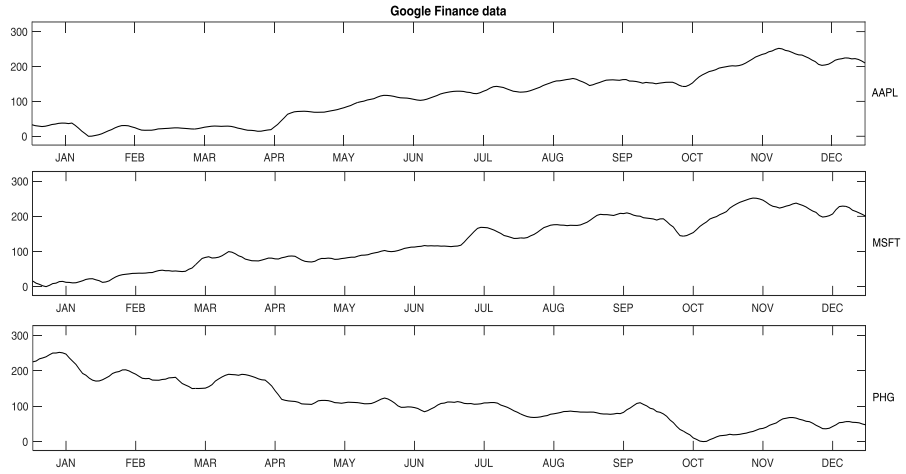

Fig. 5. Datos de 2014.

6 Resultados

En la fig. 6 se muestra el agrupamiento de series de tiempo de ocho empresas petroleras descargadas de *Google Finance*. Para cada par de series se obtienen sus patrones positivos y negativos. Se crea una nueva lista donde se cuenta cuántos patrones de longitud 1, cuántos de longitud 2, hasta la longitud máxima. Se obtiene el 15% de la longitud máxima y todos los valores por debajo de ese porcentaje se eliminan. Algunos pares quedarán sin patrones después de la eliminación pues ninguno de sus patrones fue lo suficientemente largo. Los pares que quedan con patrones son los que se imprimen en el grafo.

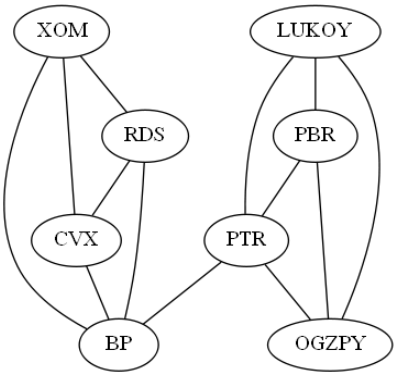

Fig. 6. Agrupamiento de empresas petroleras de países pertenecientes y no pertenecientes a la OCDE. Datos de 2014.

La asociación de patrones de empresas petroleras encontrada que casi completamente separaba en dos clústeres fuertemente conectados, uno con algunos integrantes de las consideradas siete hermanas (cuyos países pertenecen a la OCDE): BP, Chevron, Royal Dutch Shell y ExxonMobil; y otro con empresas petroleras de países no pertenecientes a la OCDE: Gazprom, Petrobras, Lukoil y Petrochina; países que tienen empresas que forman las consideradas nuevas siete hermanas. Los grupos formados están fuertemente conectados y casi completamente separados uno del otro.

Se utilizan datos de Google Finance y separan en datos de 2014 y 2015 como se aprecia en la fig. 7 y fig. 8 respectivamente. Por cada año son aproximadamente 250 puntos. Las empresas de T.I. elegidas por la disponibilidad de sus datos se muestran en la Tabla 1:

Tabla 1. Empresas analizadas en 2014 y en 2015. Elegidas por la disponibilidad de sus datos.

Tecnologías de la información	
NASDAQ:AAPL	Apple Inc.
NASDAQ:AMZN	Amazon.com, Inc.
NYSE:IBM	International Business Machines Corp.
NASDAQ:INTC	Intel Corporation
OTCMKTS:LNVGY	Lenovo Group Ltd.
NASDAQ:MSFT	Microsoft Corp.
OTCMKTS:PCRFY	Panasonic Corp.
NYSE:PHG	Koninklijke Philips NV
NYSE:SNE	Sony Corp.
NASDAQ:FB	Facebook Inc.
NASDAQ:GOOGL	Alphabet Inc.

Petróleo	
NYSE:BP	BP plc
NYSE:CVX	Chevron Corporation
OTCMKTS:OGZPY	Gazprom PAO
NYSE:PTR	PetroChina Company Limited
NYSE:RDS.A	Royal Dutch Shell plc
NYSE:XOM	Exxon Mobil Corp.
NYSE:PBR	Petroleo Brasileiro SA
OTCMKTS:LUKOY	NK LUKOIL PAO

Las series de tiempo que se analizan en el presente trabajo son series financieras, con longitud de 250 puntos aproximadamente y no se requiere reducción de dimensiones. Proponemos una medida de similitud entre series que están alineadas en tiempo, de no ser así la transformada MAT y posterior comparación no se podría llevar a cabo. La alineación obedece a la misma naturaleza de los datos obtenidos de *Google Finance* [8], sería interesante si no se asumiera tal alineación y se aplicara la transformada MAT para encontrar la LCSS, por ejemplo, pero no es el propósito del artículo. El tratamiento que se le da a la serie antes de aplicar la comparación es la transformación MAT.

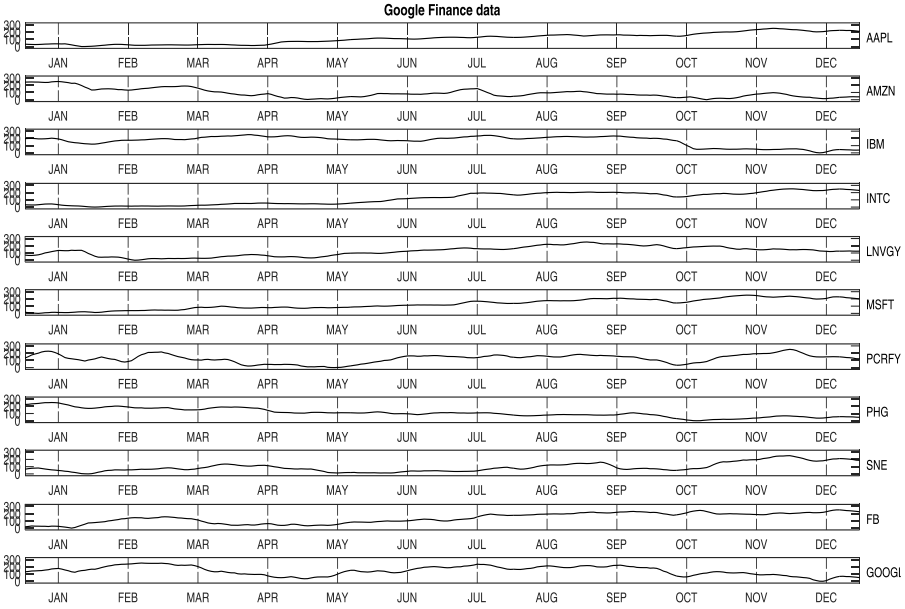

Fig. 7. Datos de TI de 2014

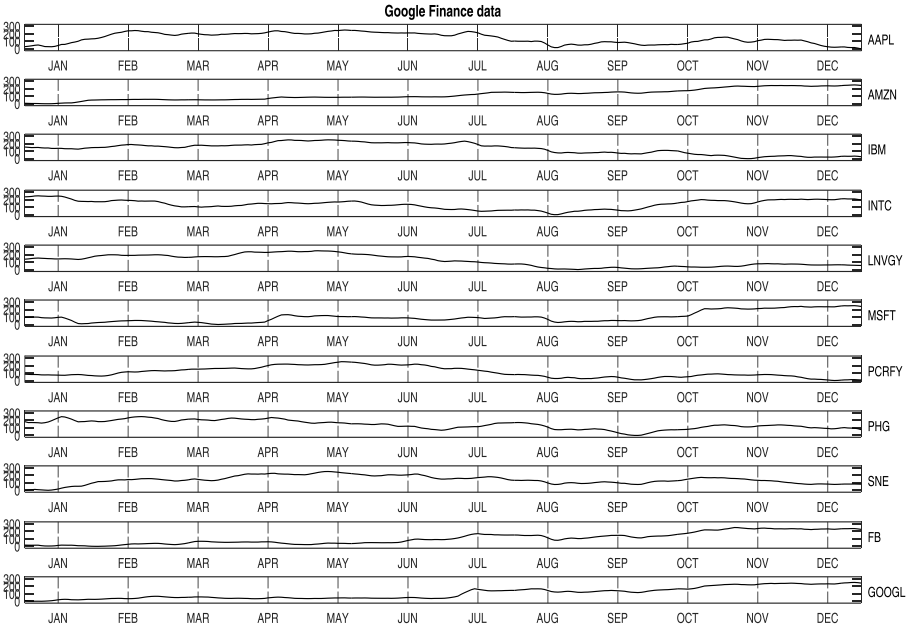

Fig. 8. Datos de TI de 2015

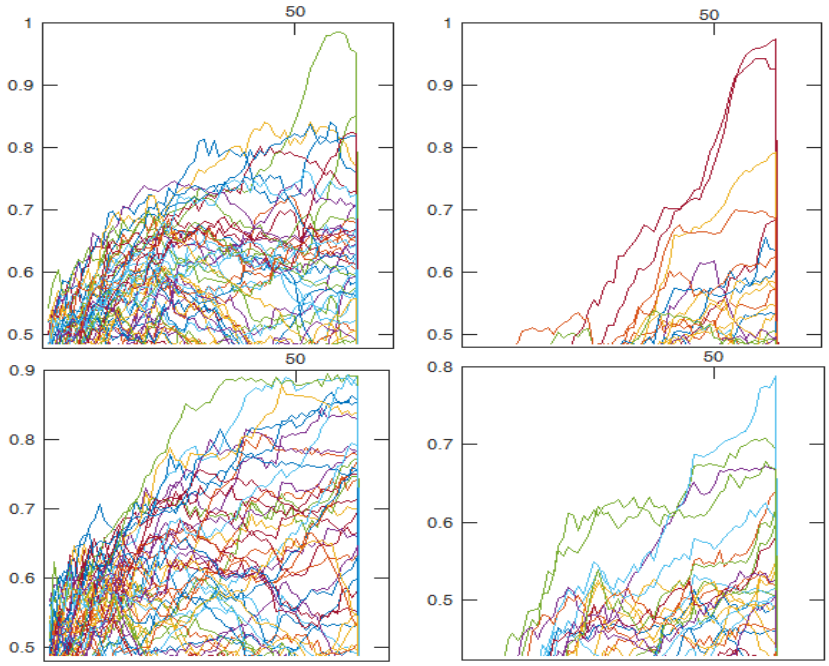

Fig. 9. Funciones de asociación para Empresas de T.I.: Arriba izquierda asociación positiva 2014, Arriba derecha asociación negativa 2014, Abajo izquierda asociación positiva 2015, Abajo derecha asociación negativa 2015

En la fig. 9 se observan las funciones de asociación, el eje *x* es el tamaño de la ventana. A continuación las ventanas sugeridas para cada gráfica. Para 2014 las ventanas sugeridas para valores positivos y las relaciones mayores a 0.8 son:

Ventana = 58: APPL-INT, APPL-MSFT, MSFT-INT, PCRFY-AMZN;
Ventana = 32: PCRFY-AMZN;
Ventana = 44: APPL-INT-LNVGY

Las ventanas sugeridas para valores negativos son de 54, 23 y 62. Pero sólo hay relaciones mayores a 0.8 para la ventana de 62: APPL-PHG, MSFT-PHG.

Para 2015 las ventanas para valores positivos sugeridas son de 39, 46 y 11; mientras que para valores negativos 62, 26 y 64. Las asociaciones positivas se dan entre (mayores a 0.8): APPL-PCRFY, PCRFY-SNE; APPL-PCRFY, PCRFY-SNE, SNE-APPL, SNE-IBM, SNE-LNVGY. Para la ventana de 11 se bajó el offset a 0.65 para mostrar la relación APPL-SNE. No hay asociaciones negativas mayores a 0.8, las más altas son con ventana de 62: IBM-GOOGL e IBM-FB, 0.70 y 0.77 respectivamente.

Una interpretación del porqué de estas relaciones no es sencilla. En el precio de una acción hay un factor de especulación. Además una empresa tiene distintas áreas y en algunas le puede ir bien como en otras mal. Las relaciones por ejemplo entre INT, APPL y MSFT vistas en 2014 se pueden interpretar como una empresa que manufactura procesadores y otras que venden los sistemas operativos. Incluso LNVGY que también

manufactura los equipos aparece en esas relaciones. Sin embargo esta explicación no se sostiene para 2015.

En 2015 existe la asociación entre PCRFY y SNE que además se puede ver desde mediados de 2014. Aunque ambos son competidores en el mercado de los televisores no son los que tienen la mayor parte del mercado (Samsung y LG). A principios de 2015 muchas empresas de televisión incluidas las cuatro mencionadas en este párrafo formaron una alianza (UHDA por sus siglas en inglés) para establecer estándares de calidad de los futuros productos de Ultra-Alta-Definición lo cual puede ser una explicación de la similitud en sus acciones.

Aunque una explicación real debería considerar el ámbito financiero, o el mercado, si sólo se consideran las series de tiempo se observa que lo que arrojan los resultados tiene sentido con respecto su forma visual.

7 Conclusión

En este artículo se propone la medición de similitud entre series de tiempo utilizando la longitud de los patrones y el número de patrones con el mismo signo. Los patrones se obtienen aplicando la transformada MAT y obteniendo la asociación de sus pendientes. A diferencia de utilizar la medida coseno, la similitud de patrones permite medir la similitud de las series de tiempo incluso cuando ésta cambia en tiempo, y además se aprovecha la similitud y disimilitud que mide la MAT.

Los trabajos anteriores [2, 7] usaban ventanas de tiempo elegidas arbitrariamente, o sea, intuitivamente. En este artículo, por primera vez se propone un método para la selección de las ventanas para obtener la información de las series de tiempo con mayor asociación entre distintas ventanas. Los valores de similitud se grafican con respecto al tamaño de la ventana, con lo cual se identifican los pares de series que son más similares para cada ventana. Los máximos de los pares de series, que además se busca que estén separados entre ellos mismos, se consideran como sugerencias para las ventanas de tiempo. Los resultados obtenidos en nuestros experimentos con datos de *Google Finance* tienen una interpretación natural tanto en el caso de las empresas petroleras, como en el caso de las empresas de tecnologías de la información.

También se propone un nuevo método para la visualización de los intervalos de tiempo mucho más clara que la que se ha usado anteriormente [7]. En el método propuesto, las asociaciones se muestran una vez que se ha seleccionado la ventana de tiempo. Este diagrama nos muestra dadas dos series de tiempo si se correlacionan positiva o negativamente en los diferentes puntos en el tiempo. Un ejemplo de la gráfica se muestra en la fig. 2, donde se puede ver que las series de las empresas CVX y XOM están muy correlacionadas en 2014, con la ventana de 30.

Los métodos que proponemos son nuevos y extienden las posibilidades de análisis series de tiempo considerados en [2, 7].

El trabajo a futuro es hacer un análisis más completo incorporando eventos de noticias que podrían ayudar en la explicación del comportamiento de series de tiempo financieras.

Agradecimientos. Este trabajo es parcialmente apoyado por los proyectos SIP 20162204 y 20161958 del Instituto Politécnico Nacional, México, y mediante beca BEIFI del mismo instituto.

Referencias

1. Esling, P., Agon, C.: Time-series data mining. ACM Computing Surveys (CSUR), Vol. 45, No. 1, p. 12 (2012)
2. Batyrshin, I., Herrera-Avelar, R., Sheremetov, L., Panova, A.: Moving approximation transform and local trend associations in time series data bases. Perception-based Data Mining and Decision Making in Economics and Finance, pp. 55–83, Springer (2007)
3. Das, G., Gunopulos, D., Mannila, H.: Time-series similarity problems and well-separated geometric sets. 13th Annual ACM Symposium on Computational Geometry. Association for Computing Machinery (1997)
4. Alcock, R.J., Manolopoulos, Y.: Time-series similarity queries employing a feature-based approach. 7th Hellenic conference on informatics, pp. 27–29 (1999)
5. Lin, J., Khade, R., Li, Y.: Rotation-invariant similarity in time series using bag-of-patterns representation. Journal of Intelligent Information Systems, Vol. 39, No. 2, pp. 287–315 (2012)
6. Ye, J., Xiao, C., Esteves, R.M., Rong, C.: Time Series Similarity Evaluation Based on Spearman's Correlation Coefficients and Distance Measures. Cloud Computing and Big Data, pp. 319–331 Springer (2015)
7. Batyrshin, I., Solovyev, V., Ivanov, V.: Time series shape association measures and local trend association patterns. Neurocomputing, Vol. 175, pp. 924–934 (2016)
8. http://www.google.com/finance