ISSN 1870-4069

Comparación cuantitativa de matrices de datos

Angel Mandujano-García, Jesús Figueroa-Nazuno, Hiram Calvo

Instituto Politécnico Nacional,
Centro de Investigación en Computación, Ciudad de México,
México

b140477@sagitario.cic.ipn.mx,{jfn,hcalvo}@cic.ipn.mx

Resumen. Son pocas las técnicas de alineamiento elástico entre matrices de datos. Las que existen realizan un análisis comparando la similitud de características, o algunas son puramente estadísticas. En este trabajo se presenta una técnica moderna para la comparación elástica de matrices numéricas que considera todos los datos disponibles de manera global y no por análisis de extracción de características como hacen otras técnicas. Se presentan resultados de la experimentación utilizando como ejemplo datos de funciones matemáticas; sin embargo, este método puede funcionar para cualquier objeto o fenómeno que pueda ser representado en forma matricial.

Palabras clave: Medida de distancia, alineamiento elástico, matrices numéricas, Fréchet moderno, funciones matemáticas.

Quantitative Comparison on Data Matrices

Abstract. There are few elastic matching techniques that take matrices as input. Perform an analysis comparing the similarity of features while the existing ones techniques are purely statistical. A modern technique for elastic matching comparison is performed in this paper. This technique considers all data in the matrices, globally, and not just analizing extracted features like other available techniques do. Experimental results are presented using a synthetic dataset of matematical functions; however this method works for any object or phenomenon that can be represented as matrix.

Keywords: Distance measurement, elastic matching, matrix representation, modern Fréchet, matematical functions.

1. Introducción

La comparación es un aspecto muy importante dentro de la vida diaria y de la computación, siempre estamos tomando decisiones que se basan en algún tipo de comparación. Cuando comparamos, lo primero que realizamos es obtener de alguna manera, las características o rasgos descriptivos de los objetos implicados en la comparación, y así, se pueden observar las diferencias y/o similitudes entre éstos. Esta estrategia en los seres humanos es realizada con naturalidad; sin embargo, dentro de la computación llevarla a cabo no es una tarea sencilla. Las computadoras necesitan de alguna herramienta matemática que se ocupe de realizar una medición cuantitativa y realizar una comparación. Con base en eso, es común realizar una métrica de los rasgos característicos de los objetos. Una alternativa, que se presenta en este artículo, es que una computadora pueda realizar una comparación no considerando rasgos descriptivos, sino tomando en cuenta todos los datos que se tienen disponibles, es decir, de manera global. La alternativa se basa en el uso de una técnica que realice un alineamiento elástico (*Elastic Matching*) sobre objetos en dos dimensiones.

Elastic Matching (EM) se ha utilizado en diversos problemas, como el reconocimiento de rostros, el reconocimiento de huellas dactilares, análisis de imágenes médicas, visión por computadora, entre otros. EM es una técnica que en general ha dado buenos resultados. Formalmente hablando, EM es definido como un problema de optimización con respecto a un mapeo elemento-elemento, lineal o no lineal [6]. En otras palabras, EM mide el esfuerzo que se tiene al ajustar A sobre B, siendo A y B dos matrices. EM optimiza el problema de alineamiento en varias dimensiones, 1D, 2D, 3D, etc. Para este trabajo se realiza un alineamiento en dos dimensiones, también llamado *Two-Dimentional Warping* (2DW). Este alineamiento ocurre en arreglos bidimensionales también llamados matrices, por lo tanto el objeto A y B son dos matrices que contienen datos numéricos. EM es un problema que presenta una complejidad computacional NP-Completo [7] debido a los grados de libertad que poseen las matrices. Las ventajas de EM sobre las técnicas de selección de rasgos son: EM es adaptativa, así generalmente posee mayor capacidad para obtener diversas deformidades que las técnicas clásicas no pueden, la optimización 2DW por sí misma, describe la deformación de carácter dominante. Este hecho muestra que EM posee propiedades útiles de técnicas de análisis estructural. EM puede estar relacionado con los marcos estadísticos y estocásticos. *Active Shape Models* y *2D HMMs* son dos buenas técnicas de ejemplos [10]. Las características de EM principalmente dependen de dos factores: (i) la formulación de 2DW, afecta el rango de deformaciones compensables. Esto quiere decir que la formulación de 2DW está relacionada con el problema que se quiere resolver; (ii) la estrategia de optimización de 2DW, afecta la precisión de los resultados de EM. De manera general, en [10] se menciona que las estrategias para obtener una solución óptima global proveen resultados más precisos que aquellas soluciones que son sub-óptimas. La técnica propuesta se basa en Programación Dinámica, donde se proporciona una solución óptima global.

Fréchet Moderno (FM) es el nombre de la técnica de comparación matricial que se ha desarrollado en el presente trabajo. FM es una medida elástica entre matrices de datos. Se le ha puesto ese nombre debido a que se basa en la Distancia de Fréchet. La distancia evaluada bajo esta aproximación es invariante a las deformaciones y surge debido a la falta de técnicas que analizan los datos representados en dos matrices de manera directa.

Existen técnicas que realizan la comparación de matrices, como la presentada en [6], que emplea una reducción de dimensionalidad en las matrices, usando Eigenvalores [9], esto es, se obtienen los datos característicos de cada matriz y eventualmente se realiza una comparación de rasgos característicos. Sin embargo son pocas y es muy difícil encontrar técnicas de comparación global. El presente trabajo se desarrolla la técnica Fréchet Moderno (FM), es una técnica muy útil cuando se requiere comparar datos de manera global, es decir, cuando se necesita considerar todos los datos que se tienen en las matrices y no los datos característicos. Se puede aplicar en diferentes ramas de la ciencia como: la minería de datos, la bioinformática, tratamiento de imágenes, reconocimiento de patrones, entre otros.

2. Antecedentes

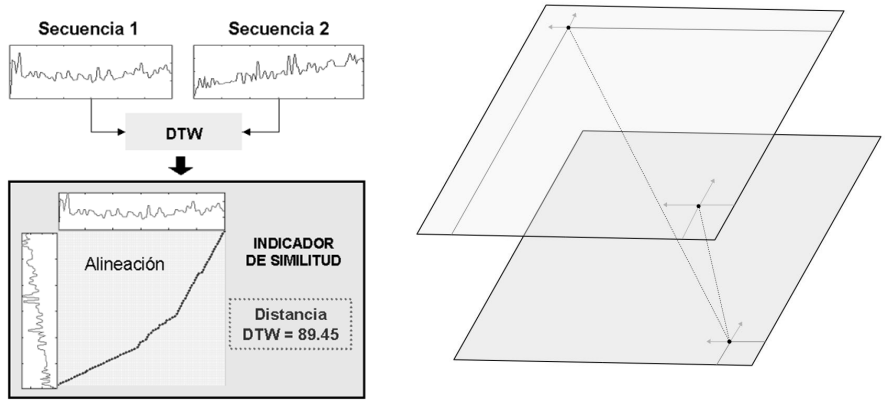
En esta sección se explican conceptos importantes para la descripción de la distancia Fréchet Moderno, se explican brevemente conceptos necesarios para la explicación y realización de la técnica:

Similitud entre matrices: El análisis convencional para la medición de similitud es mediante la selección de aquellas propiedades que se consideran importantes o relevantes; sin embargo, hacer una selección apropiada es una tarea muy difícil, aunada a esto, siempre hay pérdida de información en los datos originales, porque existe algún proceso de descarte [4]. El reto de investigación es desarrollar procedimientos sin hacer dicha selección, sino utilizar técnicas que permitan emplear toda la información disponible sobre las entidades involucradas. Para este caso se trata de matrices de datos numéricos.

Distancia de Fréchet: Maurice Fréchet fue un matemático francés que hizo trabajos de topología muy importantes. En su tesis doctoral [5, 8], presenta un procedimiento general para medir la similitud entre dos curvas F y G donde se tiene en cuenta el orden de los puntos a lo largo de las dos curvas o secuencias. De manera informal, Fréchet presentó una analogía sencilla: la Distancia de Fréchet es la longitud mínima de la correa necesaria para conectar a una persona con su perro que salen de paseo, ambos caminan sobre trayectorias diferentes, pero que tienen la misma dirección, sin poder regresar [1].

Dynamic Time Warping (DTW): Es una transformación que permite la expansión y comparación de una secuencia. Generalmente donde más se utiliza es en series de tiempo para la alineación local y global con respecto a otra secuencia,

con el objeto de minimizar la distancia base, que comúnmente es la distancia euclidiana. DTW es propia para computar la semejanza entre secuencias que se encuentran desfasadas una con respecto a la otra, o cuando una ellas presenta o ausenta segmentos con respecto a la otra [2]. Dadas dos secuencias F y G, DTW forma una matriz de dos dimensiones. Intuitivamente cada celda de la matriz representa un mapeo de un valor en la secuencia F con un valor de la secuencia G. El mapeo es la suma de la distancia individual de los mapeos previos antes calculados (ver Fig. 1a). La salida de DTW es la distancia mínima entre las secuencias y se encuentra en la última posición de la matriz generada. DTW está basada en la idea fundamental de la Distancia de Fréchet.



(a) Funcionamiento de *Dynamic Time Warping*.

(b) Recorrido completo de la distancia de Fréchet Moderno.

Fig. 1. Distancias en secuencias y distancias en matrices.

3. Descripción de la técnica Fréchet moderno

Fréchet Moderno (FM) extiende el método de DTW a dos dimensiones, es decir, los datos ahora no son secuencias o arreglos unidimensionales, sino matrices numéricas. FM se basa en DTW de tal manera que se pueden analizar datos en R^2. La entrada para FM son matrices de datos A y B de tamaño (P, Q) y (R, S) respectivamente. Fréchet Moderno es un método de comparación elástica que se obtiene comparando la distorsión que se encuentra en filas y columnas de la matriz A contra todos las filas y columnas de la matriz B, de esta manera se hace un recorrido completo en las dos matrices de entrada, como se muestra en la Fig. 1b.

Se forma una matriz de distancias acumuladas de cuatro dimensiones llamada M, cada celda de M se refiere a una alineación entre algunas celdas de A y B.

Esto es análogo a lo que ocurrió con DTW (ahora no son valores de secuencias, sino secuencias completas). $M(p, q, r, s)$ se refiere a una alineación entre $A(p, q)$ y $B(r, s)$.

De manera similar que en DTW, ahora se debe ir llenando la matriz M que guarda la distancia acumulada; una celda de M está dada por la mínima distancia de los mapeos previos de la posición actual. De este modo, la distancia entre las matrices de datos se encuentra en la última posición de M. Dados 4 indices (i,j,k,l), en la matriz M se calcula lo siguiente:

```
M(i,j,k,l) = min{ETAPAS_PREVIAS(i,j,k,l) +  Costo(i,j,k,l)}
costo = DTW(R1, R2)  +  DTW(C1, C2).
```

Para una sola matriz, cada coordenada (i, j) tiene tres posibles etapas previas, sujetas a las limitaciones de frontera: $(i-1, j-1)$, $(i, j-1)$ y $(i-1, j)$. En la Fig. 2 se muestran las etapas previas de las matrices separadas, es decir, los incisos $a)$, $b)$, $c)$ son las etapas previas de la coordenada (i, j) en la matriz A que sería el inciso $d)$, mientras que los incisos $e)$, $f)$, $g)$ son las correspondientes en la matriz B de una coordenada especifica $h)$.

Las etapas previas totales que se ocupan en Fréchet Moderno son el resultado de combinar las etapas previas de la matriz A con las de la matriz B, exceptuando las coordenadas $d)$ y $h)$, que indican la etapa actual que se está calculando. En otras palabras, las etapas previas son los puntos adyacentes con coordenada individual más pequeña en la matriz M. En la Tabla 1, se tienen las etapas previas totales con su respectiva función costo.

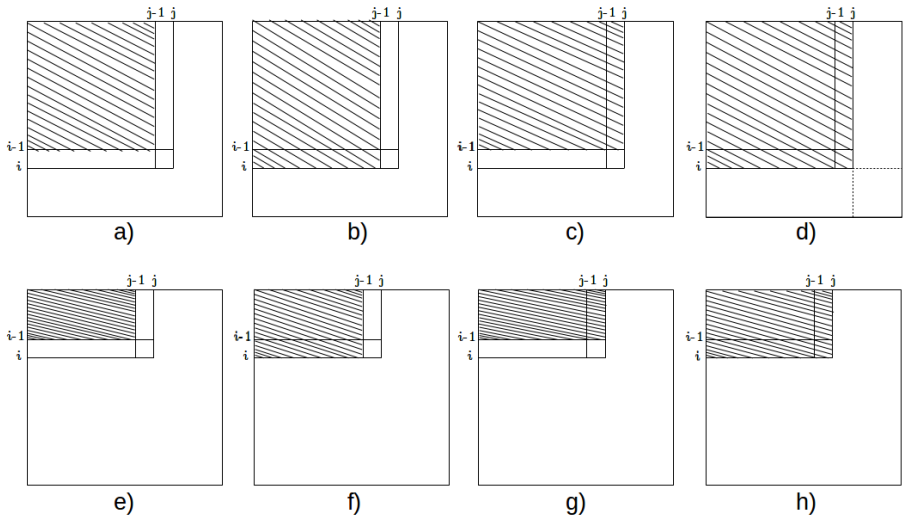


Fig. 2. Etapas previas para la matriz A, son $a)$, $b)$ y $c)$. $d)$ es la etapa actual. Similarmente son las etapas previas de la matriz B.

Tabla 1. Etapas totales con el cálculo de la función costo correspondiente a cada etapa.

#	Etapa Previa	Costo
1	$(p-1, q-1, r-1, s-1)$	$DTW(R1, R2) + DTW(C1, C2)$
2	$(p-1, q-1, r-1, s)$	$DTW(R1, R2) + DTW(C1, C2)$
3	$(p-1, q-1, r, s-1)$	$DTW(R1, R2) + DTW(C1, C2)$
4	$(p-1, q-1, r, s)$	$DTW(R1, R2) + DTW(C1, C2)$
5	$(p-1, q, r-1, s-1)$	$DTW(R1, R2) + DTW(C1, C2)$
6	$(p-1, q, r-1, s)$	$DTW(R1, R2)$
7	$(p-1, q, r, s-1)$	$DTW(R1, R2) + DTW(C1, C2)$
8	$(p-1, q, r, s)$	$DTW(R1, R2)$
9	$(p, q-1, r-1, s-1)$	$DTW(R1, R2) + DTW(C1, C2)$
10	$(p-1, q-1, r-1, s)$	$DTW(R1, R2) + DTW(C1, C2)$
11	$(p, q-1, r, s-1)$	$DTW(C1, C2)$
12	$(p, q-1, r, s)$	$DTW(C1, C2)$
13	$(p, q, r-1, s-1)$	$DTW(R1, R2) + DTW(C1, C2)$
14	$(p, q, r-1, s)$	$DTW(R1, R2)$
15	$(p, q, r, s-1)$	$DTW(C1, C2)$

La función costo consiste en utilizar DTW sobre los índices en filas y columnas de A y B. En la Tabla 1, $R1$ es la fila en la coordenada (p, q), los valores de las casillas que se encuentran entre la posición 0 a la columna q, sobre la fila p. $C1$ es la columna en la coordenada (p, q), los valores de las casillas que están entre la posición 0 a la fila p, sobre la columna q. De manera similar se obtienen $R2$ y $C2$. Ver Fig. 3. Para determinar el valor de la casilla $M(p, q, r, s)$ se debe obtener un costo para cada una de las etapas previas, que está dado por la alineación de las filas y columnas correspondientes en A y B, es decir, dada la posición $M(p, q, r, s)$ se obtienen $R1$ y $R2$, así como $C1$ y $C2$.

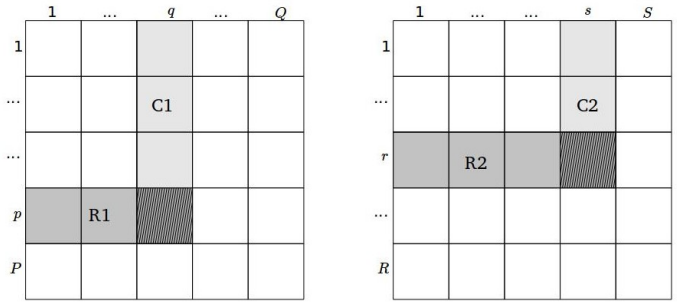


Fig. 3. Casillas que seleccionadas de cada matriz A y B respectivamente.

El algoritmo 1 es utilizado para calcular cada una de las etapas previas de la Tabla 1. La función costo se puede ir obteniendo conforme avanza la

iteración de las etapas actuales. Este proceso debe ser realizado para cada una de las coordenadas de M. Una manera eficiente de realizar este procedimiento es mediante el uso de Programación Dinámica; se debe hacer un recorrido de alta complejidad que es representado en el algoritmo.

Algoritmo 1 Algoritmo para calcular la distancia Fréchet moderno

Entrada: Matrices $A(p, q)$, $B(r, s)$ de datos numéricos.
Salida: Número escalar k, indica la distancia entre A y B.

$P \leftarrow$ Filas de A
$Q \leftarrow$ Columnas de A
$R \leftarrow$ Filas de B
$S \leftarrow$ Columnas de B
$d(P, Q, R, S)$ Matriz de 4 índices, inicializada en ∞
para $p = 2$ hasta P **hacer**
  para $q = 2$ hasta Q **hacer**
    para $r = 2$ hasta R **hacer**
      para $s = 2$ hasta S **hacer**

$$d(p,q,r,s) = min \left\{ \begin{array}{rcl} DTW(R1,R2) + DTW(C1,C2) + & d(p-1,q-1,r-1,s-1) \\ DTW(R1,R2) + DTW(C1,C2) + & d(p-1,q-1,r-1,s) \\ DTW(R1,R2) + DTW(C1,C2) + & d(p-1,q-1,r,s-1) \\ DTW(R1,R2) + DTW(C1,C2) + & d(p-1,q-1,r,s) \\ DTW(R1,R2) + DTW(C1,C2) + & d(p-1,q,r-1,s-1) \\ DTW(R1,R2) + & d(p-1,q,r-1,s) \\ DTW(R1,R2) + DTW(C1,C2) + & d(p-1,q,r,s-1) \\ DTW(R1,R2) + & d(p-1,q,r,s) \\ DTW(R1,R2) + DTW(C1,C2) + & d(p,q-1,r-1,s-1) \\ DTW(R1,R2) + DTW(C1,C2) + & d(p-1,q-1,r-1,s) \\ DTW(C1,C2) + & d(p,q-1,r,s-1) \\ DTW(C1,C2) + & d(p,q-1,r,s) \\ DTW(R1,R2) + DTW(C1,C2) + & d(p,q,r-1,s-1) \\ DTW(R1,R2) + & d(p,q,r-1,s) \\ DTW(C1,C2) + & d(p,q,r,s-1) \end{array} \right\}$$

      fin para
    fin para
  fin para
fin para
devolver $k = d(P, Q, R, S)$

4. Desarrollo

La técnica recibe como entrada dos arreglos bidimensionales de la misma clase, en este caso se utilizan matrices de datos de funciones matematicas previamente creadas y se obtiene como resultado un indicador de distancia que representa la diferencia entre los objetos de entrada, éste es un número escalar. El método de comparación responde a las propiedades de métrica [3].

Se presenta el procedimiento que se utilizó para la comparación matricial de datos numéricos. Éste es aplicado a datos generados artificialmente a partir de funciones matemáticas, pero también puede ser aplicada a cualquier fenómeno que se pueda representar en matrices, como por ejemplo grafos, ontologías, entre otros.

Creamos un *corpus* de matrices de datos numéricos provenientes de funciones matemáticas, se toma una función y se varía un valor t de manera incremental, así se produce una matriz por cada valor de t que es variado. Se repite el proceso para

las funciones que se desean comparar. Posteriormente se realizan experimentos de Búsqueda, donde se escoge una matriz a buscar (*query*) y se proporciona un directorio con varias matrices. En el experimento, nuestro algoritmo obtiene las matrices con distancia mínima con respecto a la matriz *query* que se escogió. Una muestra del conjunto de matrices se muestra en la Fig. 4. La muestra contiene tres categorías y cada categoría tiene cuatro matrices.

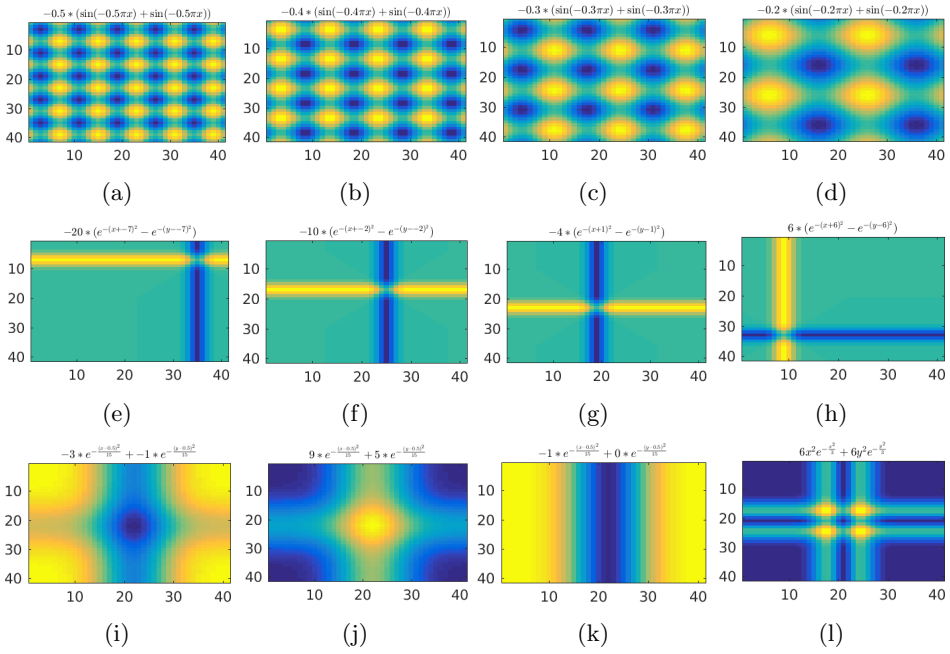


Fig. 4. Algunas matrices del corpus.

Las matrices fueron creadas por cada función; entonces es posible agruparlas como si fueran categorías. Se utilizó la técnica de distancia Fréchet Moderno para realizar la experimentación: primero se realizaron comparaciones entre elementos de las mismas categorías, posteriormente se realizaron comparaciones haciendo una mezcla de diferentes categorías. De estas matrices se selecciona una que será el *query*:

1. Se realiza la comparación dentro de la misma categoría que la del *query*, las funciones evaluadas para la categoría están en la Tabla 2, se observa el valor de la constante es variado un decimal. Los resultados se muestran en la Fig. 5, vemos la función que corresponde al *query*: $[-0.2(\sin(-0.2\pi x) + \sin(-0.2\pi x))]$, las mas cercanas son las que tienen el valor pequeño en las constantes de las funciones evaluadas.

Tabla 2. Matrices de la misma función.

ID	Cercanas
$[ID\_9]$	$-0.2 * (\sin(-0.2\pi x) + \sin(-0.2\pi x))$
$[ID\_8]$	$-0.3 * (\sin(-0.3\pi x) + \sin(-0.3\pi x))$
$[ID\_7]$	$-0.4 * (\sin(-0.4\pi x) + \sin(-0.4\pi x))$
$[ID\_6]$	$-0.5 * (\sin(-0.5\pi x) + \sin(-0.5\pi x))$

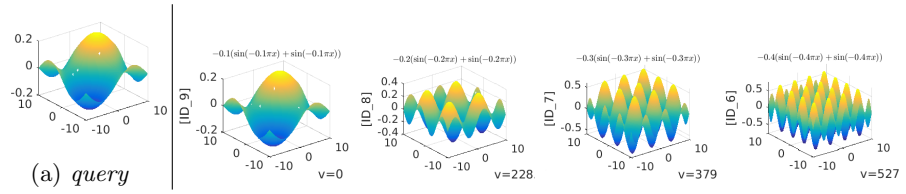


Fig. 5. Resultados de la búsqueda utilizando Fréchet Moderno. Se muestran las matrices con las distancias más cercanas al *query*, dentro de la misma categoría.

2. Se realiza una comparación considerando matrices de otras categorías. La Fig. 6 muestra en la primera columna la matriz *query* seleccionada, mientras que las otras tres columnas son las que corresponden a las matrices más cercanas. Se muestra en la parte superior la función que se evaluó para obtener esa matriz. Se puede observar que las matrices cercanas corresponden a las matrices de la misma categoría del *query*.

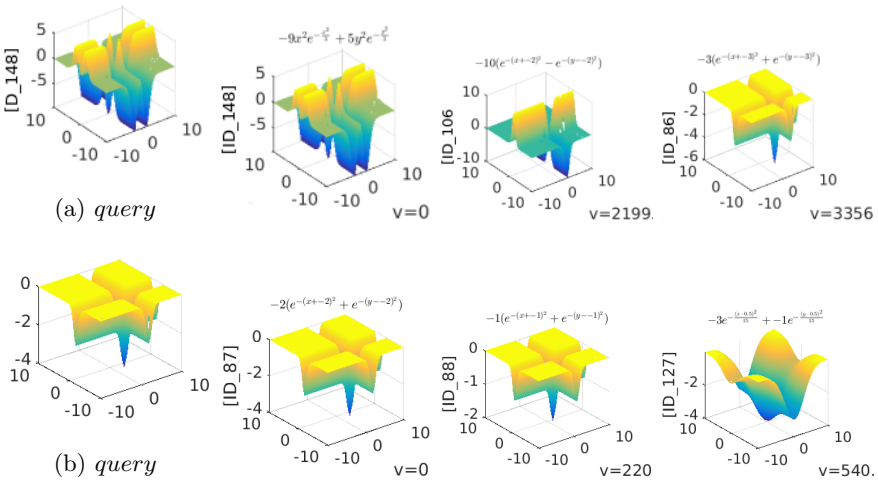


Fig. 6. Experimento donde se muestra la matriz *query*, se realiza la búsqueda en el directorio de matrices.

5. Resultados y conclusiones

Ordenando de menor a mayor el indicador que se obtuvo del método Fréchet Moderno, observamos que los números más pequeños son aquellos que están más cercanos (desde el punto de vista técnico) a la matriz *query*, es decir, las matrices más similares; de esta manera el *query* comparado contra ella misma da como resultado un cero, lo cual indica que son la misma matriz y no hay diferencia. Esto corresponde a las propiedades de las métricas.

La técnica propuesta de este trabajo es una analogía a la distancia de Fréchet, donde Fréchet ocupa una correa para conectar a un perro con su amo, y cada uno hace un recorrido. De este modo la Distancia de Fréchet Moderno también realiza un recorrido, pero es a través de datos en R^2, donde también se tiene una "correa" en cada celda visitada y va midiendo la diferencia en cada una como se mostró en la Fig. 1b. Fréchet Moderno ha resultado una técnica para la comparación de matrices de datos numéricos que es sencilla en comparación con la técnica utilizada en [6], pero es poderosa en el sentido de que considera todos los datos de las matrices. Es importante mencionar que esta técnica puede ser utilizada en muchas más situaciones, se pueden realizar experimentos con imágenes, grafos, ontologías, etc.

Referencias

1. Agarwal, P.K., Rinat, A.B., Kaplan, H., Sharir, M.: Computing the Discrete Fréchet Distance in Subquadratic Time. CoRR abs/1204.5, pp. 1–18 (2012)
2. Angeles-Yreta, A., Figueroa-Nazuno, J., Ramírez-Amaro, K.: Búsqueda de Semejanza entre Objetos 3D por Indexado. In: Reunión de Otoño Comunicación, Computacion Electrónica y Exposición Industrial, ROC&C, IEEE Sección México, Acapulco, Guerr. pp. 112–117 (2005)
3. Deza, M.M., Deza, E.: Encyclopedia of Distances, vol. 3. Springer, 3 edn. (2009)
4. Figueroa-Nazuno, J., Angeles-Yreta, A., Medina-Apodaca, J., Ortega-González, V., Ramírez-Amaro, K., Mirón-Bernal, M., Landassuri-Moreno, V.: Sobre el problema de Semejanza. Tech. rep., Centro de Investigación en Computación. Instituto Politécnico Nacional Unidad Profesional "Adolfo López Mateos", Zacatenco, México, DF (2008)
5. Fréchet, M.M.: Sur quelques points du calcul fonctionnel. Rend. del Circ. Mat. di Palermo 22(1), pp. 1–72 (1906)
6. González-Ortega, E.: Una Técnica para el Análisis de Similitud entre Imágenes. Master's thesis, Centro de Investigación en Computación, Instituto Politécnico Nacional, México (2013)
7. Keysers, D., Unger, W.: Elastic image matching is NP-complete. Pattern Recognition. Letters. 24(1-3), pp. 445–453 (2003)
8. Pitcher, A.D., Chittenden, E.W.: On the Foundations of the Calcul Fonctionnel of Fréchet. Trans. Am. Math. Soc. 19(1), pp. 66–78 (1918)
9. Rico-Martínez, J.: Determinación Numérica de Eigenvalores y Eigenvectores. Tech. rep., ICT 224, Departamento de Ingeniería Mecánica, Universidad de Guanajuato, F.I.M.E.E., Mexico (2003)
10. Uchida, S., Sakoe, H.: A survey of elastic matching techniques for handwritten character recognition. IEICE Trans. Inf. Syst. E88-D(8), pp. 1781–1790 (2005)