LA ESTRUCTURA DE BURKHARD-KELLER EN LA BUSQUEDA DE LAS CADENAS MAS SIMILARES A UN CONJUNTO SOBRE EL QUE EXISTE DEFINIDA UNA DISTRIBUCION DE PROBABILIDAD

AUTORES: SANTANA, O.; PEREZ, J.; HERNADEZ, Z.; RODRIGUEZ, A.
Departamento de Informática y Sistemas
Universidad de Las Palmas de Gran Canaria
Apto.: 550. Las Palmas de Gran Canaria. España.

RESUMEN:

En este trabajo se plantea el problema de la búsqueda de las cadenas más similares a un conjunto de cadenas sobre el que existe una distribución de probabilidad que expresa la fiabilidad con la que cada una de las cadenas representa a la cadena en cuestión. El concepto de similitud es en el sentido de Levenshtein, [LE66].

0.- INTRODUCCION:

En este trabajo se plantea el problema de la búsqueda de las cadenas más similares, de entre las de un Diccionario dado constituido por cadenas de caracteres que pertenecen a un alfabeto, a una cadena dada que se encuentra definida por medio de un conjunto de cadenas y una distribución de probabilidad asociada que expresa la fiabilidad con la que cada una de las cadenas del conjunto representa a la cadena incógnita en cuestión. El concepto de similitud es en el sentido de Levenshtein, [LE66], utilizado posteriormente por Wagner y Fisher, [WF74]. En la sección 1 se generalizan la Distancia de Levenshtein y la Distancia Invariante Transposicional, [SD87], a un conjunto de cadenas, con una distribución de probabilidad asociada; y se estudian sus propiedades. En la sección 2 se plantean dos esquemas de búsqueda, uno con evolución del radio de búsqueda decreciente y otro creciente, sobre la estructura de Burkhard-Keller organizada según la Distancia Invariante Transposicional [SP88], [SP89a], [SP89b] y [SP90]. En la sección 3 se comentan las distribuciones de probabilidad sobre los conjuntos de cadenas. En la sección 4 se presentan los resultados experimentales y las conclusiones.

1.- DISTANCIAS ENTRE UNA CADENA Y UN CONJUNTO:

Sea $\underline{C}=\{X_1,X_2,...,X_n\}$ un conjunto de cadenas de caracteres sobre un alfabeto dado; sean $p_1,p_2,...,p_n$ números reales mayores que cero que sumen uno, o sea, una distribución de probabilidad sobre $\underline{C}$; y sea X una cadena de caracteres sobre el mismo alfabeto.

Se define la Distancia de Levenshtein entre $\underline{C}$ y X, $\underline{DL}(\underline{C},X)$, así:

$$\underline{DL}(\underline{C},X)=\sum_{i=1}^{n} p_i*DL(X_i,X)$$

donde $DL(X_i,X)$ es la Distancia de edición de Levenshtein entre la cadenas X_i y X, [LE66] [WF74] y [LV85a/b].

Análogamente se define la Distancia Invariante Transposicional entre $\underline{C}$ y X, $\underline{DIT}(\underline{C},X)$, como:

$$\underline{DIT}(\underline{C},X)=\sum_{i=1}^{n} p_i*DIT(X_i,X)$$

donde $DIT(X_i,X)$ es la Distancia Invariante Transposicional entre las cadenas X_i y X, [SD87].

1.1- Propiedades:

Para cualquier par de cadenas **X** e **Y** y para todo conjunto $\underline{\boldsymbol{C}}$, se verifica que:

i) $\underline{\boldsymbol{DL}}(\underline{\boldsymbol{C}},\mathbf{X}) \geq \mathbf{0}$.

ii) Si $\underline{\boldsymbol{DL}}(\underline{\boldsymbol{C}},\mathbf{X})=\mathbf{0}$ entonces $\underline{\boldsymbol{C}}=\{\mathbf{X}\}$.

iii) Si $\underline{\boldsymbol{C}}=\{\mathbf{Y}\}$ entonces $\underline{\boldsymbol{DL}}(\underline{\boldsymbol{C}},\mathbf{X})=\mathbf{DL(Y,X)}$.

iv) $\underline{\boldsymbol{DL}}(\underline{\boldsymbol{C}},\mathbf{X}) + \mathbf{DL(X,Y)} \geq \underline{\boldsymbol{DL}}(\underline{\boldsymbol{C}},\mathbf{Y})$.

v) $\underline{\boldsymbol{DIT}}(\underline{\boldsymbol{C}},\mathbf{X}) \leq \underline{\boldsymbol{DL}}(\underline{\boldsymbol{C}},\mathbf{X})$.

De estas propiedades se deduce que esta distancia es una extensión a conjuntos de cadenas de la Distancia de Levenshtein clásica entre cadenas de caracteres y que para conjuntos de una sola cadena coincide con ésta.

i') $\underline{\boldsymbol{DIT}}(\underline{\boldsymbol{C}},\mathbf{X}) \geq \mathbf{0}$.

ii') Si $\underline{\boldsymbol{DIT}}(\underline{\boldsymbol{C}},\mathbf{X})=\mathbf{0}$ entonces $\underline{\boldsymbol{C}}$ está formado por sinónimos transposicionales de **X**.

iii') Si $\underline{\boldsymbol{C}}=\{\mathbf{Y}\}$ entonces $\underline{\boldsymbol{DIT}}(\underline{\boldsymbol{C}},\mathbf{X})=\mathbf{DIT(Y,X)}$.

iv') $\underline{\boldsymbol{DIT}}(\underline{\boldsymbol{C}},\mathbf{X})+\mathbf{DIT(X,Y)}\geq\underline{\boldsymbol{DIT}}(\underline{\boldsymbol{C}},\mathbf{Y})$.

De estas propiedades se deduce que esta distancia es una extensión a conjuntos de cadenas de la Distancia Invariante Transposicional entre cadenas de caracteres y se cumplen propiedades análogas.

DEMOSTRACION DE LAS PROPIEDADES

i) $\underline{DL}(\underline{C},X)\geq 0$ ($\underline{DIT}(\underline{C},X)\geq 0$ para **i'**)) ya que es una combinación lineal con coeficientes positivos de cantidades no negativas.

ii) Si $\underline{DL}(\underline{C},X)=0$ entonces $\forall i\in\{1,...,n\}$: $DL(X_i,X)=0$ (por ser $\underline{DL}$ una combinación lineal con coeficientes positivos de DLs no negativas) y por tanto $X_i=X$ $\forall i\in\{1,...,n\}$, con lo que $\underline{C}=\{X\}$.

ii') Si $\underline{DIT}(\underline{C},X)=0$ entonces $\forall i\in\{1,...,n\}$: $DIT(X_i,X)=0$ (por ser $\underline{DIT}$ una combinación lineal con coeficientes positivos de DITs no negativas) y por tanto X_i es sinónimo transposicional de X $\forall i\in\{1,...,n\}$.

iii) Si $\underline{C}=\{Y\}$ la única distribución de probabilidad posible es con un sólo valor p=1 y por tanto $\underline{DL}(\underline{C},X)=DL(Y,X)$ ($\underline{DIT}(\underline{C},X)=DIT(Y,X)$ para **iii')**).

iv) $$\underline{DL}(\underline{C},X) + DL(Y,X) = \sum_{i=1}^{n} p_i{*}DL(X_i,X) + DL(Y,X) =$$

$$= \sum_{i=1}^{n} p_i{*}DL(X_i,X) + \sum_{i=1}^{n} p_i{*}DL(Y,X) = \sum_{i=1}^{n} p_i{*}[DL(X_i,X)+DL(Y,X)] \geq$$

$$\geq \sum_{i=1}^{n} p_i{*}DL(X_i,Y) = \underline{DL}(\underline{C},Y)$$

La propiedad que se aplica es la desigualdad triangular para la DL entre cadenas.

Para **iv')** se aplica una argumentación paralela sobre DIT.

v) $\underline{DIT}(\underline{C},X) = \sum_{i=1}^{n} p_i * DIT(X_i,X) \leq \sum_{i=1}^{n} p_i * DL(X_i,X) = \underline{DL}(\underline{C},X)$

1.2.- Cota superior de *DL*:

Dado un conjunto de cadenas $\underline{C}=\{X_1,X_2,...,X_n\}$ con una distribución de probabilidad asociada que expresa la fiabilidad con la que cada una de las cadenas representa a la cadena incógnita **X**, $p_1,p_2,...,p_n$. Llámese π_i a la distorsión de X_i, es decir, $\pi_i=DL(X_i,X)/|X|$, para cada $i\in\{1,...,n\}$, donde $|X|$ es la longitud de la cadena **X**; y sea $\pi=max\{\pi_i/1\leq i\leq n\}$.

Supóngase que $\forall i\in\{1,...,n\}$: X_i se obtiene de **X** a través de I_i inserciones, E_i extracciones y S_i sustituciones, en número mínimo, se tiene que: $DL(X_i,X)=I_i+E_i+S_i$.

Despejando E_i: $E_i=DL(X_i,X)-I_i-S_i$

como: $I_i\geq 0$, $E_i\geq 0$ y $S_i\geq 0$,

se tiene que: $E_i\leq DL(X_i,X)$,

por otra parte: $|X|=|X_i|+E_i-I_i$,

luego:

$$\begin{aligned} DL(X_i,X) = \pi_i * |X| &= \pi_i * (|X_i|+E_i-I_i) \leq \\ &\leq \pi_i * (|X_i|+E_i) \leq \\ &\leq \pi_i * (|X_i|+DL(X_i,X)) \leq \\ &\leq \pi * (|X_i|+DL(X_i,X)) \end{aligned}$$

y multiplicando por p_i y sumando desde **1** hasta **n** se obtiene:

$$\underline{DL}(\underline{C},X) \leq \pi * (\sum_{i=1}^{n} p_i * |X_i| + \underline{DL}(\underline{C},X))$$

de donde se obtiene que:

$$\underline{DL}(\underline{C},X) \leq (\sum^{n} p_i * |X_i|) * \pi/(1-\pi)$$

$$i=1$$

De lo que se infiere que si las cadenas del conjunto de búsqueda no han sufrido distorsiones superiores a π, con respecto a la cadena que representan, el segundo miembro de la expresión anterior constituye una cota superior de la ***DL*** entre el conjunto y la cadena.

2.- ESQUEMAS DE BUSQUEDA:

Dados un alfabeto de caracteres sobre el que estarán definidas todas las cadenas, una base de cadenas que se denominará Diccionario y un conjunto de cadenas $\underline{C}=\{X_1,X_2,...,X_n\}$ (pertenecientes o no al Diccionario) con una distribución de probabilidad asociada, $p_1,p_2,...,p_n$. El problema que se plantea es encontrar el subconjunto de las cadenas del Diccionario, **CMS(DLM)**, que se encuentran a mínima Distancia de Levenshtein, ***DL***, del conjunto $\underline{C}$.

El Diccionario se encuentra estructurado en forma de árbol de Burkhard-Keller, organizado según la Distancia Invariante Transposicional, **BK-DIT**, [SP88].

2.1.- Esquema decreciente:

El esquema de búsqueda decreciente para conjuntos de cadenas, **BK-DIT+*DL*-D**, está basado en el esquema de búsqueda en el árbol de Burkhard-Keller, [SP88]. El radio de búsqueda, **DLM**, se inicializa a un valor que es cota superior de la ***DL*** mínima entre $\underline{C}$ y las cadenas del Diccionario y decrece a medida que

se encuentren cadenas a menor distancia ***DL***.

Al acceder a un nodo se calcula ***DIT*****(*C*,W)**, donde **W** es la cadena alojada en el nodo. Si ***DIT*****(*C*,W)>DLM**, se continúan explorando todos aquellos ramales comprendidos entre el redondeo a entero superior de ***DIT*****(*C*,W)-DLM** y el redondeo a entero inferior de ***DIT*****(*C*,W)+DLM**, accediendo a los ramales por proximidad al valor de ***DIT*****(*C*,W)**. En el resto de los ramales las cadenas se encuentran a una ***DIT*** de ***C*** superior a **DLM**, en virtud de **iv')**, y, por **v)**, también a una ***DL*** superior a **DLM**. Si ***DIT*****(*C*,W)≤DLM** se calcula ***DL*****(*C*,W)**. Si ***DL*****(*C*,W)<DLM** se actualiza el valor de **DLM**, **DLM=*DL*(*C*,W)**, y el conjunto de más similares, **CMS(DLM)**, contiene a **W** como única cadena; si ***DL*****(*C*,W)=DLM** se añade **W** a **CMS(DLM)**.

2.2.- Esquema creciente:

El esquema de búsqueda que se planteará aquí tiene una evolución del radio de búsqueda, **DLM**, creciente, **BK-DIT+*DL*-C**, inspirado en los esquemas crecientes expuestos en [SP90]. De esta forma su valor inicial será cero e irá creciendo de uno en uno al completar cada uno de los recorridos a través del árbol, hasta que se encuentre respuesta.

Al acceder a cada nodo se procede igual que en el esquema decreciente, en cuanto a las acciones a tomar respecto a ***DIT*** y ***DL***. Se almacenan los valores de las distancias calculadas en la propia estructura a fin de que en los recorridos posteriores a través de la misma puedan ser leídos y no tengan que ser evaluados de nuevo. El orden en el que se exploran los ramales, en este esquema, no es relevante, salvo para los dos ramales más extremos.

3.- DISTRIBUCIONES DE PROBABILIDAD SOBRE LOS CONJUNTOS:

En la introducción se ha mencionado que cada conjunto de búsqueda tendrá una distribución de probabilidad asociada que expresa la fiabilidad con la que cada una de las cadenas del conjunto representa a la cadena incógnita. Existen muchas posibles distribuciones y deberán estar relacionadas con la Distancia de edición de Levenshtein entre cada una de las cadenas y la cadena incógnita.

En los procesos de simulación experimental se tratan conjuntos de cadenas procedentes de una misma cadena incógnita que han sido alteradas por un distorsionador que introduce errores de edición al azar en cuanto a la posición y al tipo de perturbación de que se trate.

A fin de poder estudiar el comportamiento de los esquemas de búsqueda frente a la cardinalidad del conjunto y a la distorsión promedio del mismo, en una primera fase se han construido conjuntos en los que todas sus cadenas se encuentran a una misma distancia de la cadena incógnita, a tales conjuntos se les asigna la distribución uniforme como probabilidad asociada.

A continuación, al objeto de estudiar la influencia de la dispersión interna del conjunto así como la de la adecuación de la distribución de probabilidad asociada, se han construido conjuntos que no se encuentran homogéneamente distribuidos en cuanto a su distancia de la cadena incógnita respectiva y a tales conjuntos se les ha asociado dos posibles distribuciones: la uniforme y una distribución, teóricamente más adecuada, en la que la probabilidad asociada es inversamente proporcional a la **DL** entre la cadena

incógnita y cada una de las cadenas del conjunto.

4.- RESULTADOS EXPERIMENTALES:

Los experimentos se han llevado a cabo sobre un diccionario con 11278 palabras en español.

El cálculo de cada una de las **DIT(X_i,W)**, necesarias para obtener la ***DIT*(*C*,W)**, se puede realizar aprovechando la compartición de componentes de **DIT** expuesta en [SP89a], sin más que ampliar el campo de almacenamiento de la **DIT** parcial a una posibilidad de almacenamiento para cada una de las posibles cadenas del conjunto de búsqueda.

Cuando se tratan conjuntos homogéneos, independientemente del esquema de búsqueda que se utilice, Figura 1, los aciertos —respuestas entre las que se encuentra la cadena

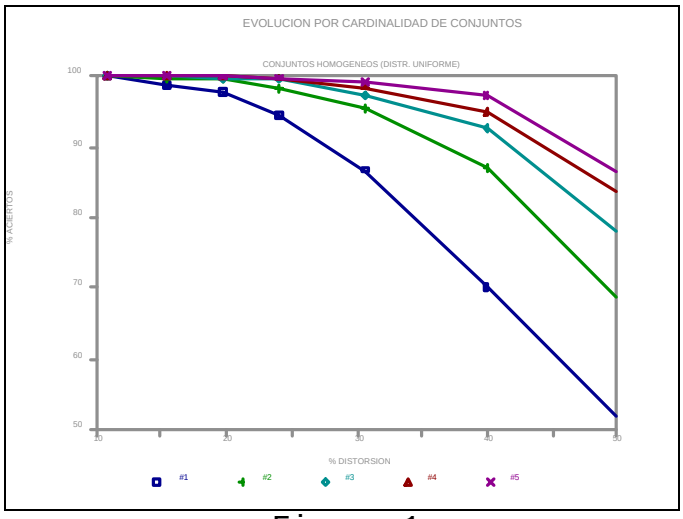

Figura 1

de la cual se ha obtenido el conjunto en cuestión por medio de la actuación del distorsionador— aumentan considerablemente a medida que se incrementa el tamaño del conjunto, hasta el punto de mantenerse razonablemente elevados (por encima del 80%) para distorsiones muy grandes (más del 40%). No sólo se obtienen más aciertos sino que además con mayor eficacia ya que las respuestas únicas se intensifican de una manera notoria y la multiplicidad de la respuesta disminuye de manera drástica, Figura 2, manteniéndose en

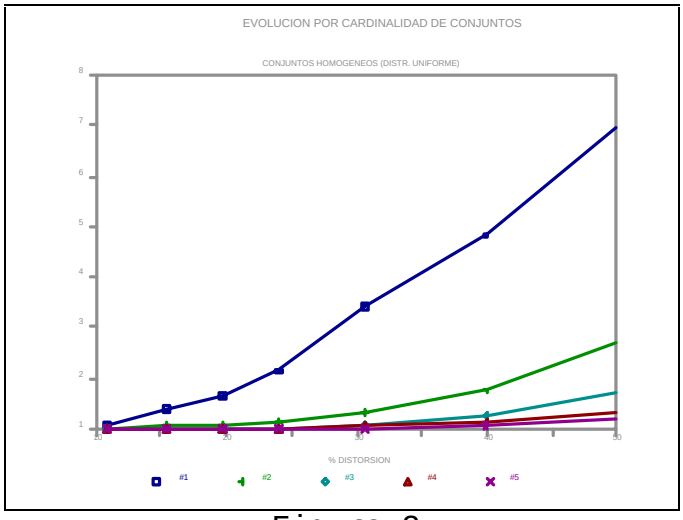

Figura 2

promedio por debajo de 3 incluso para conjuntos de dos cadenas y para las distorsiones más elevadas.

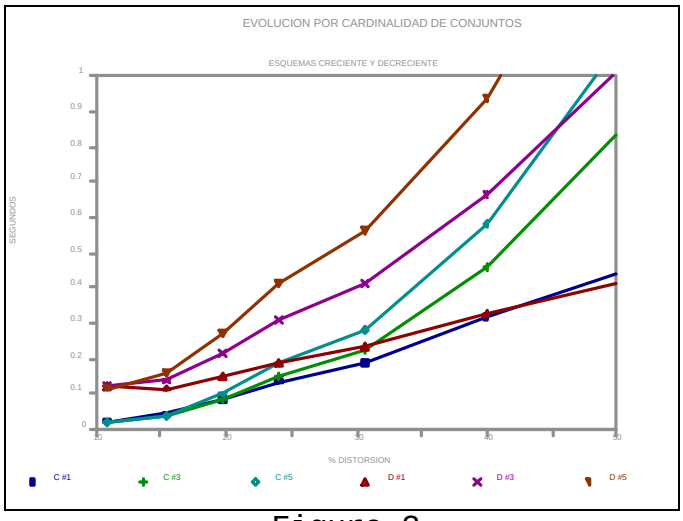

Figura 3

La realización de los esquemas de búsqueda, Figura 3, empeora con el aumento del tamaño de los conjuntos y con la distorsión, siendo mejor el esquema creciente que el decreciente para un mismo tamaño de conjunto, excepto para conjuntos de tamaño uno y distorsiones superiores al 40%.

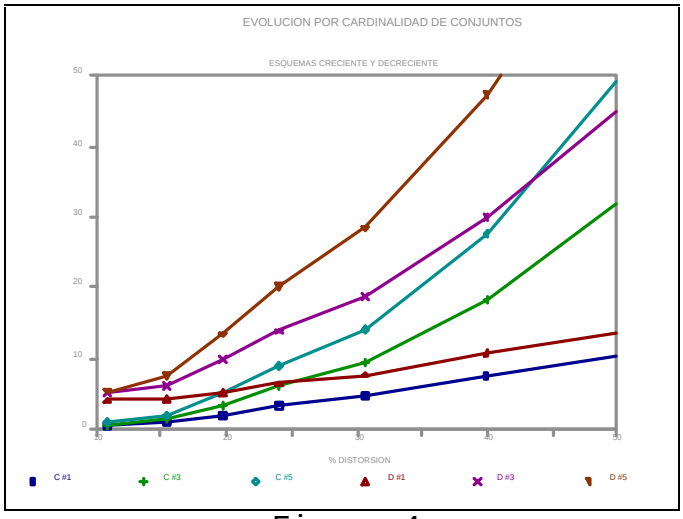

Figura 4

El comportamiento de estos esquemas es imputable fundamentalmente, Figura 4, al costo de los cálculos de las distancias implicadas, aunque debe tenerse en cuenta que para los esquemas crecientes existe un mayor costo imputable al recorrido por la estructura a medida que aumenta la distorsión, Figura 5, llegando este efecto a sobrepasar el menor costo computacional en distancias para conjuntos de tamaño uno, como se ha comentado anteriormente.

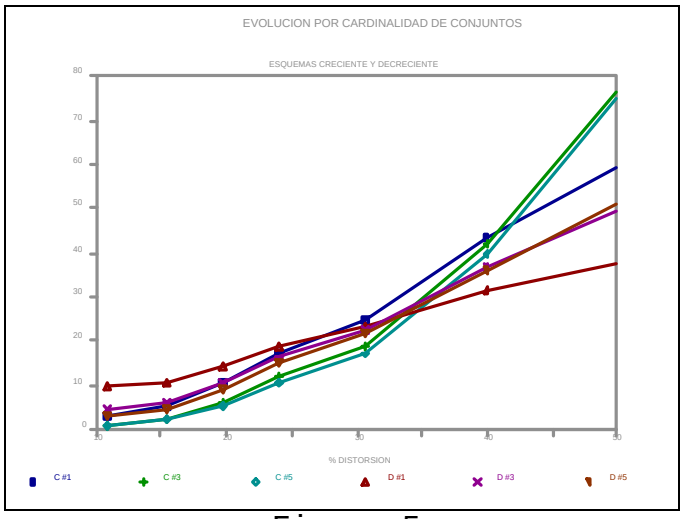

Figura 5

En las pruebas realizadas con conjuntos no homogéneos se han seleccionado los conjuntos de tamaño 3 para poner de manifiesto la influencia de la distribución utilizada, para los demás tamaños de conjunto los resultados resultan similares.

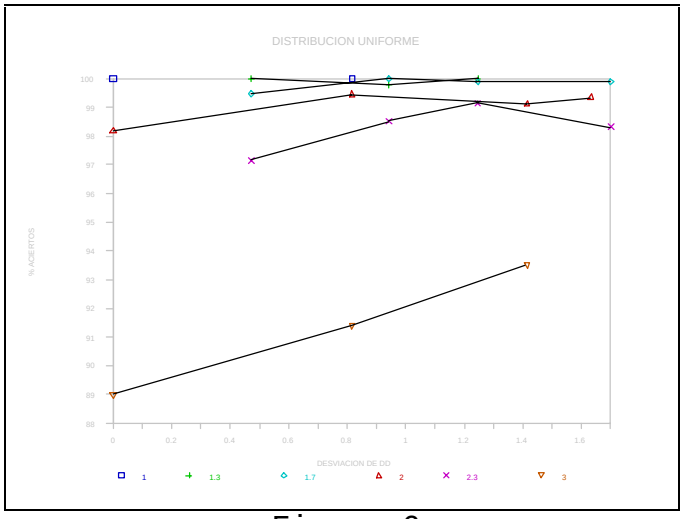

Figura 6

En general se observa, Figuras 6 y 7, un mayor número de aciertos cuando la distribución es la adecuada. Esto pone de manifiesto el hecho de que la distribución debe representar fielmente la fiabilidad con que cada una de las cadenas representa a la cadena incógnita. Ocurre además que cuando la distribución es adecuada, resulta favorable el hecho de que exista una mayor dispersión de los valores de las distancias.

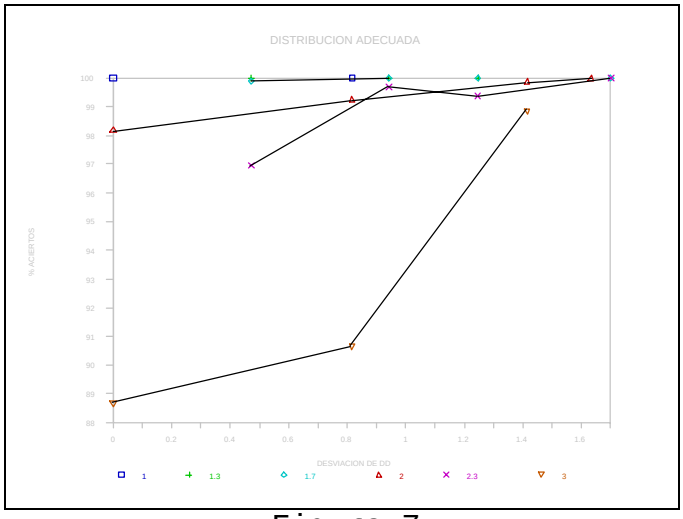

Figura 7

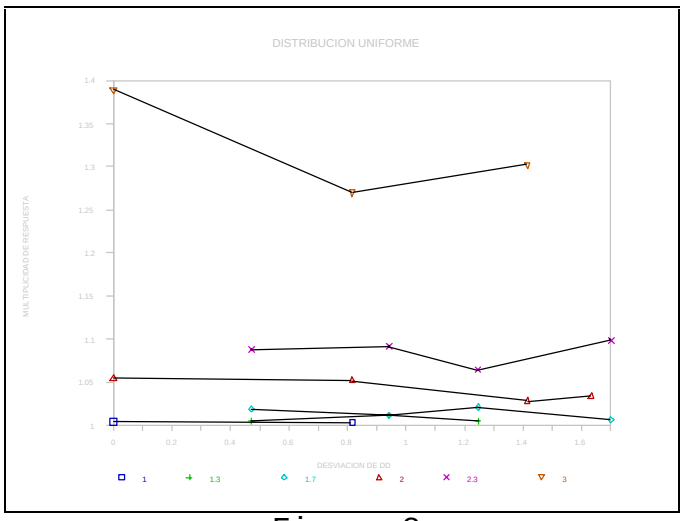

Figura 8

Con las multiplicidades de las respuestas, Figuras 8 y 9, ocurre un efecto similar, en el sentido de que la distribución adecuada favorece la eficacia y este efecto mejora cuando la dispersión de las distancias en el conjunto aumenta. Se mantiene que a mayores valores medios de distancia se producen peores respuestas, de forma análoga al caso homogéneo. La respuesta única sigue pautas similares.

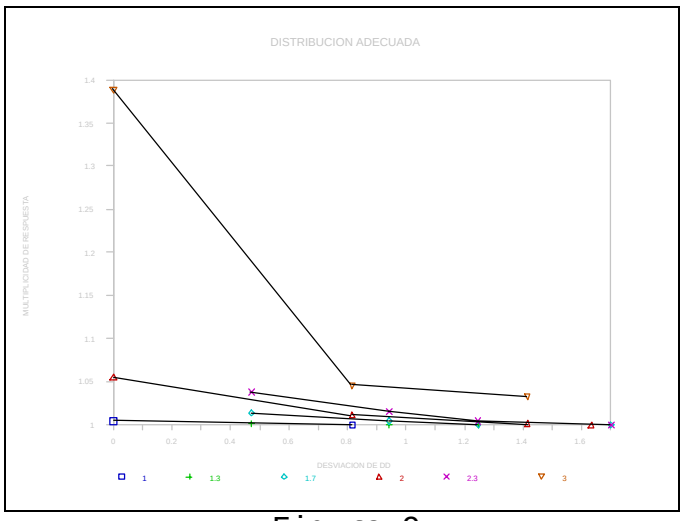

Figura 9

Los esquemas de búsqueda creciente y decreciente se comportan de manera semejante, Figuras 10, 11, 12 y 13, aunque la realización del creciente es mejor que la del decreciente. Si la distribución utilizada es la uniforme, los esquemas de búsqueda son insensibles a la dispersión de las distancias, pero si la distribución es adecuada mejoran su realización a medida que se tiene una mayor dispersión. En todo caso persiste el hecho de que a medida que la distancia promedio crece la realización empeora.

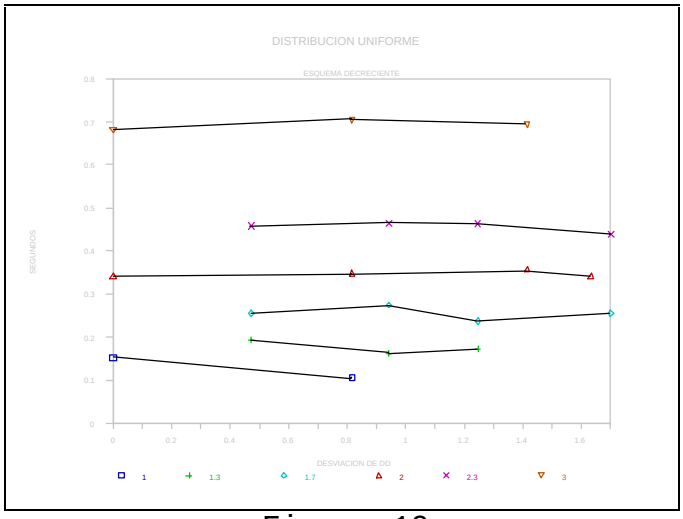

Figura 10

Como conclusión se constata que el uso de conjuntos, con distribución de probabilidad asociada, para la búsqueda, aporta una información mayor que una única cadena y como consecuencia se obtienen mejores resultados, esto es, más aciertos, menor multiplicidad en la respuesta y una mayor frecuencia de respuestas únicas. Además, la distribución asociada debe expresar adecuadamente

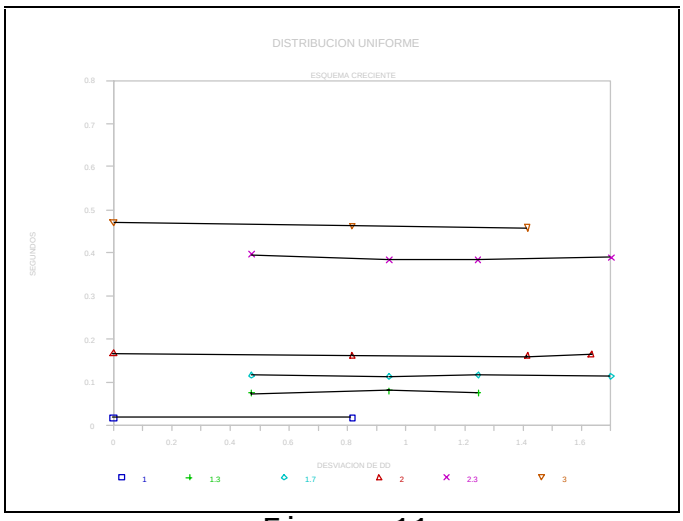

Figura 11

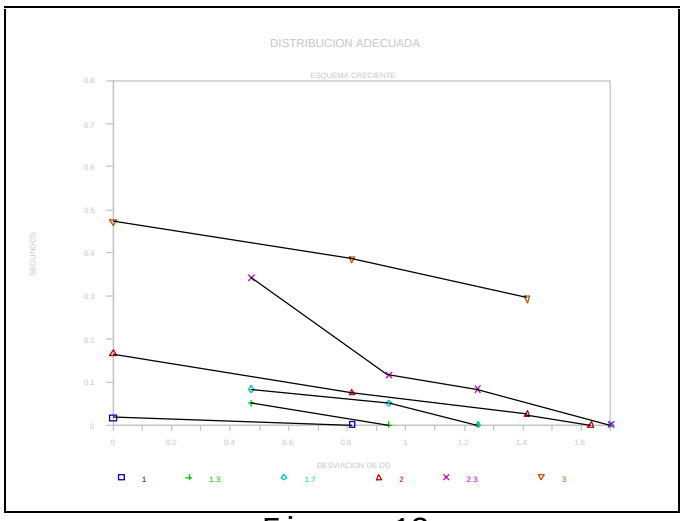

Figura 13

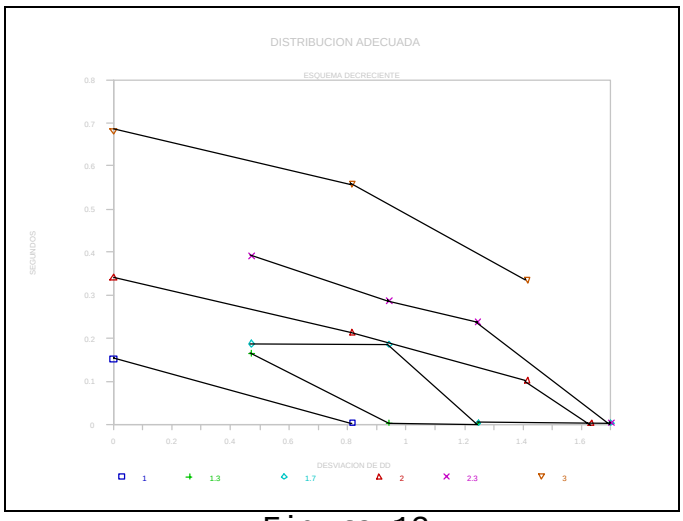

Figura 12

la fiabilidad con la que cada una de las cadenas representa a la cadena incógnita, en cuyo caso resulta favorable el hecho de que exista dispersión entre los valores de distancia dentro del conjunto. En general es más eficaz el esquema creciente que el decreciente. La mayor complejidad computacional inherente al cálculo de las distancias a conjuntos hace que los procesos de búsqueda empeoren su realización a medida que la cardinalidad de los mismos aumenta, a pesar de que se acceda a una porción más reducida del Diccionario.

BIBLIOGRAFIA:

[LE66] LEVENSHTEIN, V.I.: "Binary Codes Capable of Correcting, Insertions and Reversals". Soviet Phys. Dokl. 10, 707/710, (1966).

[LV85a] LANDAU, G.M.; VISHKIN, U.: "Efficient String Matching in the Presence of Errors". Proc. 26th IEEE FOCS, 126/136, (1985).

[LV85b] LANDAU, G.M.; VISHKIN, U.: "Efficient String Matching with k Differences". TR-36/85, Department of Computer Science, Tel Aviv University, Submitted for Journal Publication, 1985.

[LV86a] LANDAU, G.M.; VISHKIN, U.: "Efficient String Matching with k Mismatches", Theoretical Computer Science, 43, 239/249, (1986).

[LV86b] LANDAU, G.M.; VISHKIN, U.; NUSSINOV, R.: "An Efficient String Matching Algorithm with k Differences for Nucleotide and Amino Acid Sequences". Nucleic Acid Research 14 (1), 31/46, (1986).

[SD87] SANTANA, O.; DIAZ, M.; MAYOR, O.; REYES, J.: "Esquemas y estructura para la búsqueda de las palabras más similares a una dada". XIII Conferencia Latinoamericana de Informática, Vol. II, 1169/1189, (1987).

[SP88] SANTANA, O.; PEREZ, J.; LOPEZ G.; RODRIGUEZ, G.: "La estructura de Burkhard-Keller en la búsqueda de las cadenas más similares a una dada". XIV Conferencia Latinoamericana de Informática, Buenos Aires, Argentina, (1988).

[SP89a] SANTANA, O.; PEREZ, J.; HERNANDEZ, Z; RODRIGUEZ H., G.: "Sharing the Components of Transposition-Invariant Distance, DIT, on DIT-organized Burkhard-Keller Structure in Searches for Best Matching Strings". IEEE INTERNATIONAL WORKSHOP ON TOOLS FOR ARTIFICIAL INTELLIGENCE "Achitectures, Languages & Algorithms", Fairfax, Virginia, U.S.A., October (1989).

[SP89b] SANTANA, O.; PEREZ, J.; ESPINO, M.; RODRIGUEZ, J.C.: "Referencias Distanciales de Levenshtein en la Estructura de Burkhard-Keller Organizada según la Distancia Invariante Transposicional. Parte I". Actas de la XV Conferencia Latinoamericana de Informática, Santiago de Chile, Julio, Vol.II, 327/334, (1989).

[SP90] SANTANA, O.; PEREZ, J.; RODRIGUEZ, J.C.: "Increasing Radius Search Schemes for the Most Similar Strings on the Burkhard-Keller Tree". Cybernetics and Systems: An International Journal, 21: 167-180, 1990.

[UK83] UKKONEN, E.: "On Approximate String Matching". Proc. Int. Conf. Found. Comp. Theor., Lecture Notes in Computer Science 158, Springer-Verlag, 487/495, (1983).

[UK85] UKKONEN, E.; "Finding Approximate Pattern in Strings". J. of Algorithms, 6, 132/137, (1985).

[WF74] WAGNER, R.A.; FISCHER, M.J.: "The String-to-String Correction Problem". JACM, 21 (1), 168/173, (1974).