ESQUEMAS Y ESTRUCTURA PARA LA BUSQUEDA DE LAS PALABRAS MAS SIMILARES A UNA DADA

O. SANTANA, M. DIAZ, O. MAYOR, J. REYES
E.U. Informática; Universidad Politécnica de Canarias
Aptdo. 550, Las Palmas de Gran Canaria, España.

En este trabajo se aborda el problema de la recuperación desde un diccionario del conjunto de palabras más similares a una palabra de búsqueda. Se introduce una distancia, DIT, invariante ante las trasposiciones, a fin de reducir el tiempo de proceso mediante la restricción del uso de la distancia direccional de Wagner y Fisher, según se muestra en el estudio experimental. Asimismo, se estudian variaciones en el esquema de búsqueda, a fin de poder desplazarse sobre el binomio exactitud-rapidez.

Seguidamente se estructuran los componentes de la DIT, produciendo con ello una disminución radical del tiempo de proceso. Se estudian las alternativas de optimización en la propia estructura, lo que conduce a nuevas reducciones del tiempo de proceso e incluso del almacenamiento utilizado por la estructura.

Aparte de las medidas estáticas tomadas durante el estudio de la estructura, se lleva a cabo un estudio experimental de su comportamiento frente al esquema de búsqueda. Al igual que en el del esquema secuencial se plantea de nuevo el dilema entre exactitud y velocidad de proceso. Por último, se introduce una optimización al esquema de búsqueda estructurado, con la consiguiente disminución del tiempo de proceso.

ESQUEMAS Y ESTRUCTURA PARA LA BUSQUEDA DE LAS PALABRAS MAS SIMILARES A UNA DADA

O. SANTANA, M. DIAZ, O. MAYOR, J. REYES
E.U. Informática; Universidad Politécnica de Canarias
Aptdo. 550, Las Palmas de Gran Canaria, España.

Resumen:

El problema que se aborda en este trabajo consiste en la recuperación desde un diccionario del conjunto de palabras más similares a una palabra de búsqueda. Se introduce el cálculo de una distancia que es independiente de la posición que ocupan los caracteres en las palabras, utilizada como un filtro para el cálculo de la distancia de Wagner y Fischer [6] a fin de mejorar el rendimiento del esquema de búsqueda cuando sólo se utiliza como criterio de similitud esta última distancia.

0.- INTRODUCCION.

Desde que surgieron los computadores, el esfuerzo investigador ha estado mucho más dirigido hacia las máquinas que hacia las personas que las usan. Uno de los factores que más complican la comunicación hombre-máquina reside en el hecho de que éstas exigen perfección en sus entradas, perfección que el hombre apenas puede darle. Los ordenadores son muy ágiles en descubrir errores, pero desgraciadamente su agilidad parece acabar ahí, su respuesta se suele limitar a un mensaje de error y una invitación para que la persona que lo maneja corrija su orden y la introduzca de nuevo. ¿Por qué los computadores han de ser tan intolerantes con los errores si son tan certeros en descubrirlos?. Cuando las personas se comunican, los errores se suceden con frecuencia, pero éstos se ignoran generalmente, permitiendo que la comunicación sea fluída. Sería sumamente interesante que los ordenadores tuvieran una tolerancia a los errores similar a la de las personas.

Alberga[2] usó matrices binarias para evaluar medidas de aproximación entre dos palabras. Szanser[4] desarrolló un proceso matemático de coincidencia elástica que era efectivo en un 95%. Morgan[3] realizaba un o-exclusivo entre dos palabras para determinar si había ocurrido un único error simple. Wagner y Fischer[6] desarrollaron un algoritmo que puede determinar la distancia entre dos palabras medida como el mínimo coste de la secuencia de operaciones de edición: sustitución, inserción y extracción.

En las secciones 1 y 2 se establecen los tipos de errores de edición que puede sufrir una palabra, se define y calcula la distancia direccional, DD (distancia de Wagner y Fischer[6]). En la sección 3 el problema se ha planteado como la recuperación secuencial a partir de un diccionario del conjunto de palabras que tengan una mínima distancia direccional a la palabra de búsqueda. En la sección 4 se define un tipo de distancia, DIT, invariante a las trasposiciones, cuyo valor es siempre menor o igual al duplo de la distancia direccional, DD. En la sección 5 el esquema de búsqueda secuencial de la sección 3 se ha optimizado debido a la inclusión del cálculo de la distancia DIT como filtro adaptivo para el cálculo de las distancias direccionales. Se propone una estructuración de los componentes que intervienen en el cálculo de DIT en la sección 6 [1,5]. En las secciones 7, 8 y 9 se estudian los parámetros más relevantes de dicha estructura, así como los criterios adoptados a fin de optimizarla. En la sección 10 se describe

el esquema de búsqueda de las palabras más similares a una dada mediante un planteamiento integrado DIT y DD recorriendo la estructura propuesta. Los tiempos promedios de búsqueda utilizando dicha estructura se presentan y se comparan en la sección 11 con los correspondientes a la recuperación secuencial. En la sección 12 se experimenta con una modificación ulterior del esquema de búsqueda propuesto, en el cual se garantiza que no existe ninguna palabra distorsionada por encima de un valor dado. La sección 13 presenta las conclusiones de este trabajo.

1.- DISTANCIA DIRECCIONAL.

1.1 ERRORES POSIBLES EN UNA PALABRA.

Wagner considera tres tipos posibles de errores básicos que pueden darse en una palabra:

1.- Sustitución de un carácter por otro.
2.- Extracción de un carácter.
3.- Inserción de un carácter.

Se considera que todas las palabras que se mencionan en lo sucesivo están compuestas por una secuencia de caracteres extraídos de un alfabeto $\{\alpha 1,..,\alpha m\}$.

Sea X una palabra, sea X<i> el carácter que está en la i_ésima posición de la palabra X; X<i:j> la secuencia de caracteres que va de X<i> a X<j> ambos inclusive, si i>j entonces $X<i:j>=\mu$, la hilera nula. |X| denota la longitud de X.

1.2 OPERACIONES DE EDICION.

Una operación de edición es el par $(\theta,\Omega)<>(\mu,\mu)$ donde θ,Ω son palabras de longitud menor o igual que 1, es decir, o son μ o son un único caracter. La palabra Y resulta de aplicar (θ,Ω) a la palabra X, que se escribe X->Y, si $X=\sigma\theta\gamma$ e $Y=\sigma\Omega\gamma$.

Dada (θ,Ω), es una operación de sustitución si $\theta<>\mu$, $\Omega<>\mu$ y $\theta<>\Omega$; es una operación de extracción si $\Omega=\mu$ y es una operación de inserción si $\theta=\mu$.

Sea S una secuencia s1, s2, ..., sn de operaciones de edición. Una derivación-S de X hasta Y es una secuencia de palabras X0, X1, ..., Xn tal que X=X0, Y=Xn y Xj-1->Xj vía sj para j=1,..,n.

Se dice que S convierte X en Y si hay una derivación-S de X hasta Y.

1.3 FUNCION DE COSTE.

Sea Γ una función arbitraria de coste que asigne a cada operación de edición (θ,Ω) un número real no negativo $\Gamma(\theta,\Omega)$. Se puede extender Γ a la secuencia S definiendo:

$$\Gamma(S)=\sum_{j=1}^{n}\Gamma(sj) \quad \text{si } n>=1 \text{ y } \Gamma(S)=0 \text{ si } n=0$$

Se llama distancia direccional de edición al mínimo coste de todas las secuencias de edición que transformen X en Y. Formalmente $\sigma(X,Y)=\min\{\Gamma(S)\}$.

2.- CALCULO DE LA DISTANCIA DIRECCIONAL.

Si X e Y son dos palabras cualesquiera, se define $X(i)=X\langle 1:i\rangle$, $Y(j)=Y\langle 1:j\rangle$ y $DD(i,j)=\sigma(X(i),Y(j))$.

$$DD(0,0)=0$$

$$DD(i,0)=\sum_{r=1}^{i}\Gamma(X\langle r\rangle,\mu) \qquad DD(0,j)=\sum_{r=1}^{j}\Gamma(\mu,Y\langle r\rangle)$$

Es decir, el coste de convertir μ en sí mismo es cero, el coste de reducir la palabra X a μ es la suma de los costes de extraer todos los caracteres de X y el coste de obtener Y a partir de μ es el la suma de los costes de añadir cada uno de los caracteres de Y.

Para calcular $DD(i,j)$ $i=1..|X|$, $j=1..|Y|$ hay que tener en cuenta tres casos.

1.- Convertir $X(i-1)$ en $Y(j-1)$ y $X\langle i\rangle$ en $Y\langle j\rangle$.
2.- Convertir $X(i-1)$ en $Y(j)$ y extraer $X\langle i\rangle$.
3.- Convertir $X(i)$ en $Y(j-1)$ y añadir $Y\langle j\rangle$.

De estas tres posibilidades se escoge aquella cuyo coste sea mínimo.

Por lo tanto:

$$DD(i,j)=\min\{DD(i-1,j-1)+\Gamma(X\langle i\rangle,Y\langle j\rangle),\ DD(i-1,j)+\Gamma(X\langle i\rangle,\mu),\ DD(i,j-1)+\Gamma(\mu,Y\langle j\rangle)\}$$

Y la distancia direccional entre X e Y viene dada por $DD(|X|,|Y|)$.

3.- CONJUNTO DE PALABRAS MAS SIMILARES USANDO SOLO DD.

Con esta distancia direccional se puede definir el conjunto de palabras similares a una palabra dada:

Sea D el conjunto de todas las palabras correctas cuyo cardinal es ND. Sea B una palabra de búsqueda, sea P un subconjunto de D definido como sigue:

$$P=\{D1,D2,\ldots,Dn/Di\in D\ i=1..n\ y\ \sigma(B,Di)=\sigma(B,Dj)=DD\ i,j=1..n\ y\ DD\ es\ mínimo\}$$

Entonces P es el conjunto de palabras que más se aproximan a B.

3.1 ESQUEMA DE BUSQUEDA DDS. RESULTADOS EXPERIMENTALES.

El esquema de búsqueda DDS realiza una búsqueda secuencial en D, determinando la DD de cada palabra con respecto a B. Si la DD es inferior a la distancia mínima actual se actualiza la distancia mínima y el conjunto de palabras más similares a B. Si es igual se actualiza P.

Si es mayor se prueba la palabra siguiente en D. Si en algún momento DD es cero entonces B está en D.

En este trabajo se supone que el coste de cualquier operación de edición es uno.

Se evaluaron el porcentaje de aciertos, el porcentaje de respuesta única, así como el promedio de multiplicidad de respuesta utilizando el esquema de búsqueda DDS con un diccionario formado por las 2089 palabras de uso más frecuente del idioma Castellano.

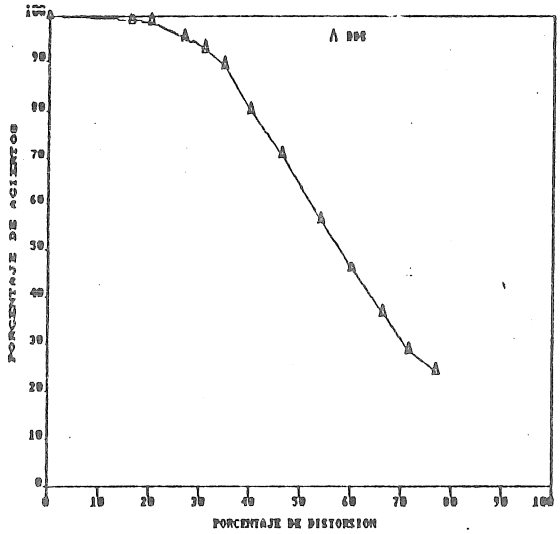


Figura 1 "Gráfica de porcentajes promedio de aciertos en el esquema de búsqueda DDS"

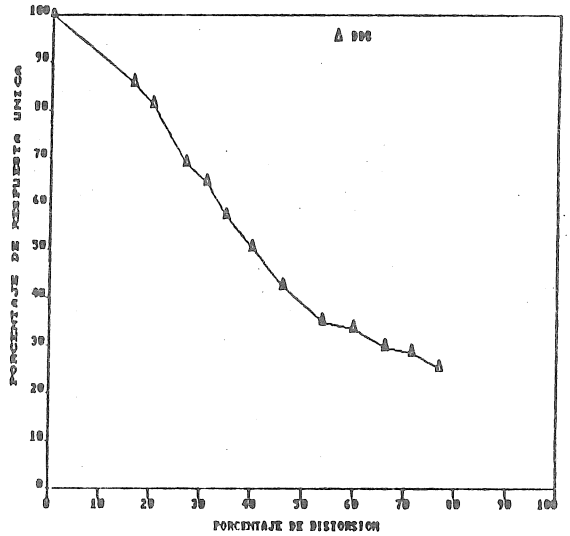


Figura 2 "Gráfica de los porcentajes promedio de respuesta única en el esquema de búsqueda DDS"

Se dice que una palabra X ha sufrido una distorsión π originando otra palabra Y si:

$$\pi = DD(X,Y)/|Y|$$

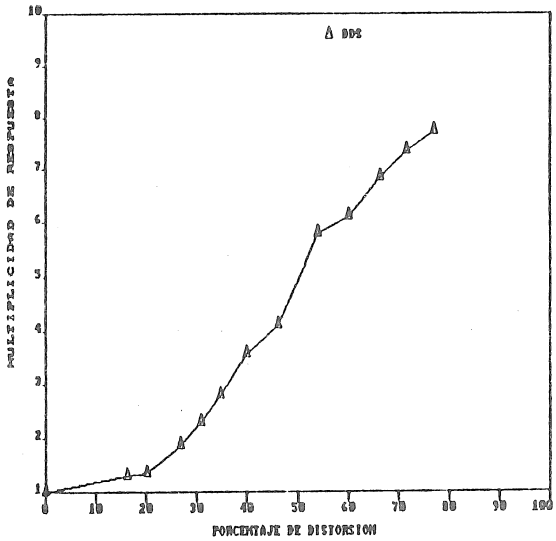


Figura 3 "Gráfica de los promedios de multiplicidad de respuesta en el esquema de búsqueda DDS".

El porcentaje de aciertos se mantiene razonablemente alto hasta distorsiones del 30%, figura 1, por allá del cual cae rápidamente. Para distorsiones del 50% la probabilidad de que el resultado sea verdadero se sitúa en torno al 0.5.

Las respuestas únicas no son demasiado abundantes incluso para distorsiones bajas, ver figura 2.

En la figura 3 se observa que el número de palabras como respuestas no-únicas aumenta rápidamente a partir del 20% de distorsión.

3.2 LIMITE SUPERIOR DE LA DD EN FUNCION DE LA DISTORSION MAXIMA PERMITIDA.

Si una palabra X ha sufrido una distorsión π originando otra palabra Y cuya DD(X,Y) se obtiene a partir de una secuencia que contiene $\Phi1$ inserciones, $\Phi2$ extracciones y $\Phi3$ sustituciones se tiene que:

$$\Phi1+\Phi2+\Phi3 = DD(X,Y)$$

despejando $\Phi2$: $\Phi2 = DD(X,Y) - \Phi1 - \Phi3$

debido a que: $\Phi1, \Phi2, \Phi3 >= 0$

se tiene que: $\Phi2 <= DD(X,Y)$

Por otra parte: $|X| = |Y| + \Phi2 - \Phi1$

$$DD(X,Y) = \pi * |X| = \pi * (|Y| + \Phi2 - \Phi1)$$
$$DD(X,Y) <= \pi * (|Y| + \Phi2)$$
$$DD(X,Y) <= \pi * (|Y| + DD(X,Y))$$

de donde:

$$DD(X,Y) <= \pi/(1-\pi) * |Y|$$

De lo que se infiere que si la palabra de búsqueda B ha sufrido una distorsión menor o igual que π la inicialización de la distancia direccional máxima es:

$$DDM = \pi/(1-\pi) * |B|$$

3.3 COSTE DE COMPUTACION Y ALMACENAMIENTO DEL ESQUEMA DE BUSQUEDA DDS.

Según Wagner la complejidad de cálculo de la DD(X,Y) entre dos palabras X e Y es $O(|X|*|Y|)$. Debido a que la búsqueda es exhaustiva, en el caso de que una palabra no se encuentre en el diccionario, la complejidad de cálculo de DDS en toda la base de datos y el coste de almacenamiento de esta estructura secuencial son respectivamente:

$$O(\ |B|* \sum_{j=1}^{ND} |Dj|\ ) \qquad O(\ \sum_{j=1}^{ND} |Dj|\ )$$

El bajo perfomance del esquema de búsqueda DDS, ver figura 4, conduce a la pregunta: ¿Es necesario el cómputo de DD(B,Dj) para todas

las palabras del diccionario D? Se verá que no.

4.- DISTANCIA INVARIANTE TRASPOSICIONAL.

4.1 DEFINICION.

Sean X e Y dos palabras, se definen xαi e yαi como el número de veces que αi esté contenido en X y en Y respectivamente. Se observa que:

$$|X| = \sum_{i=1}^{m} x\alpha i$$

Donde m es el número de caracteres que componen el alfabeto.

Se define la distancia invariante trasposicional DIT como:

$$DIT(X,Y) = \sum_{i=1}^{m} [abs(x\alpha i - y\alpha i)] + abs(|X| - |Y|)$$

Hay que tener en cuenta que si bien en el cálculo de la DD no hay nada que se pueda tener precalculado ya que todo el cálculo depende tanto de B como de Dj. En el cómputo de DIT el vector de frecuencias de caracteres de B es absolutamente independiente de la correspondiente Dj con la que se esté comparando por lo que se puede calcular sólo una vez al principio del proceso. Si además junto a cada palabra de la base de datos se almacena su vector de frecuencias de caracteres, entonces la complejidad de cálculo de DIT secuencial, DITS, es:

$$O(ND*m)$$

y el coste de almacenamiento será:

$$O(ND*m + \sum_{j=1}^{ND} |Dj|)$$

4.2 COMPORTAMIENTO DE DIT RESPECTO A DD.

Obsérvese como se comporta la distancia DIT ante sustituciones, inserciones y extracciones.

Sea X una palabra que se convierte en Y1 mediante Φ1 inserciones, Φ2 extracciones, Φ3 sustituciones. Sea C el subconjunto de los caracteres de X e Y que no han sido afectados por ninguna operación de edición, C1 el subconjunto de caracteres afectados por las inserciones, C2 el subconjunto de caracteres afectados por extracciones y C3 el

subconjunto de caracteres afectados por sustituciones. Supóngase por simplicidad que:

$$C1 \cap C2 = C1 \cap C3 = C2 \cap C3 = \phi$$

$$\sum_{\alpha h \in C} abs(x\alpha h - y1\alpha h) = 0 \qquad \sum_{\alpha i \in C1} abs(x\alpha i - y1\alpha i) = \Phi 1$$

$$\sum_{\alpha j \in C2} abs(x\alpha j - y1\alpha j) = \Phi 2 \qquad \sum_{\alpha k \in C3} abs(x\alpha k - y1\alpha k) = 2*\Phi 3$$

$$abs(|X|-|Y1|) = abs(\Phi 1 - \Phi 2) <= \Phi 1 + \Phi 2$$

$$DIT = \sum_{\alpha h \in C} abs(x\alpha h - y1\alpha h) + \sum_{\alpha i \in C1} abs(x\alpha i - y1\alpha i) +$$

$$+ \sum_{\alpha j \in C2} abs(x\alpha j - y1\alpha j) + \sum_{\alpha k \in C3} abs(x\alpha k - y1\alpha k) +$$

$$+ abs(|X|-|Y1|) <= 0 + \Phi 1 + \Phi 2 + 2*\Phi 3 + \Phi 1 + \Phi 2 = 2*(\Phi 1 + \Phi 2 + \Phi 3)$$

En este caso $DD(X,Y1) = \Phi 1 + \Phi 2 + \Phi 3$ por lo que:

$$DD(X,Y) >= DIT(X,Y)/2$$

La DIT mayor se obtiene al imponer la condición de disjunción. Como en el caso, ya estudiado, de múltiples sustituciones, si la disjunción entre los conjuntos Ci no se cumple, puede suceder que la interacción entre errores de distinto tipo que afecten al mismo carácter haga que estos influyan en la DIT en menor medida que la suma de cada uno de ellos por separado, por lo tanto:

$$DIT = \sum_{i=1}^{m} abs(x\alpha i - y1\alpha i) + abs(|X|-|Y1|) <=$$

$$<= \Phi 1 + \Phi 2 + 2*\Phi 3 + abs(\Phi 1 - \Phi 2) <= \Phi 1 + \Phi 2 + 2*\Phi 3 + \Phi 1 + \Phi 2 = 2*\Phi 1 + 2*\Phi 2 + 2*\Phi 3$$

En este caso $DD(X,Y1) = \Phi 1 + \Phi 2 + \Phi 3$ por lo que, en general:

$$DD(X,Y) >= DIT(X,Y)/2$$

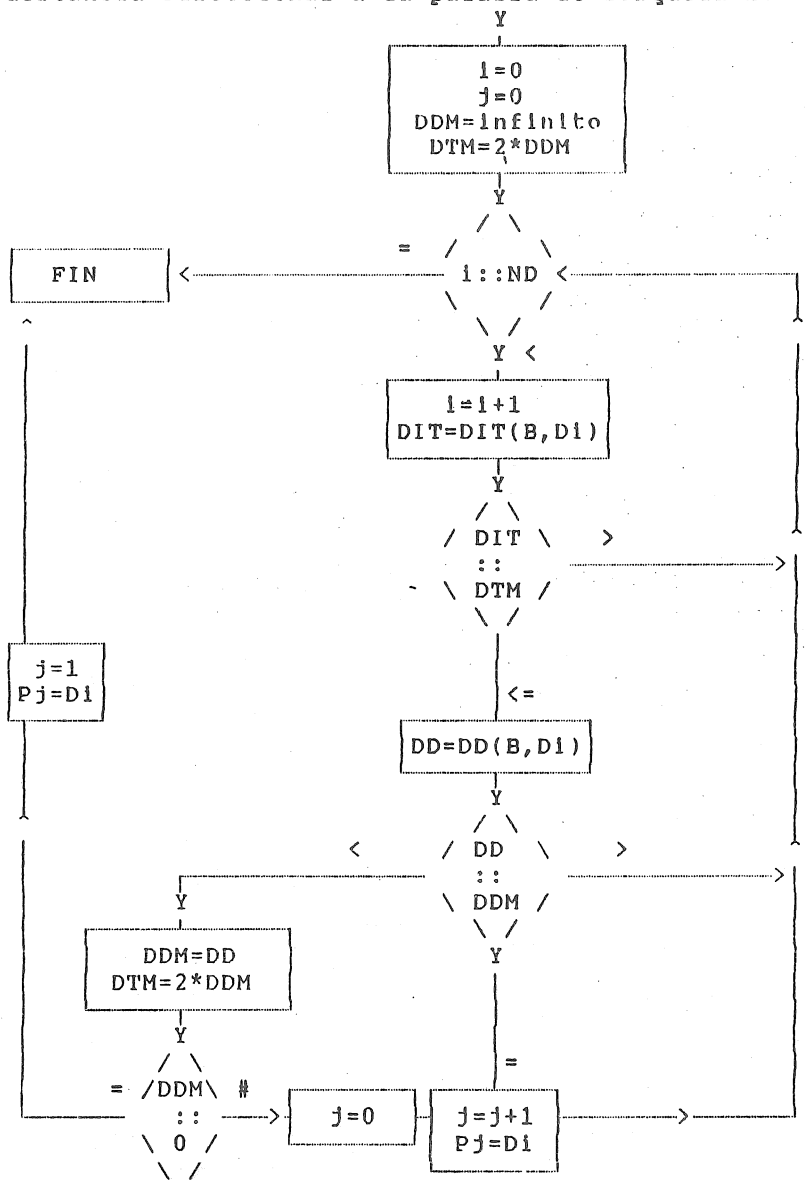
5.- CONJUNTO DE PALABRAS MAS SIMILARES USANDO DIT COMO FILTRO ADAPTIVO PARA EL CALCULO DE DD.

5.1 ESQUEMA DE BUSQUEDA DITS+DD.

De lo anterior surge la idea de disponer de un umbral DTM y calcular DD sólo para aquellas palabras cuyo DIT<=DTM. Además DTM puede ser ajustado dinámicamente. Teniendo en cuenta la relación:

$$DIT(B,Dj) <= 2*DD(B,Dj)$$

toda palabra Dj tal que su distancia DIT a la palabra de búsqueda B sea mayor que dos veces la distancia direccional mínima actual no puede tener una DD menor o igual a ésta y por tanto no se ha de calcular su distancia direccional a la palabra de búsqueda B.



Es destacable que este proceso tiene un ciclo de realimentación que le permite adaptar el filtro según la distorsión sufrida por la palabra de búsqueda.

5.2 COSTE DE COMPUTACION Y ALMACENAMIENTO DEL ESQUEMA DE BUSQUEDA DITS+DD.

El cálculo de DITS+DD presenta la siguiente complejidad:

$$O(ND*m+|B|*\sum_{j\in J}|Dj|)$$

siendo J={j/Dj verifica DIT(B,Dj)<=DTM ajustable}

El coste de almacenamiento es

$$O(ND*m+\sum_{j=1}^{ND}|Dj|)$$

5.3 RESULTADOS EXPERIMENTALES DEL ESQUEMA DE BUSQUEDA DITS+DD.

La Distancia Direccional Máxima se ha inicializado a:

1) infinito

2) |B|/2. A fin de mejorar los tiempos a costa de perder exactitud para palabras de búsqueda distorsionadas más de un 33.33%.

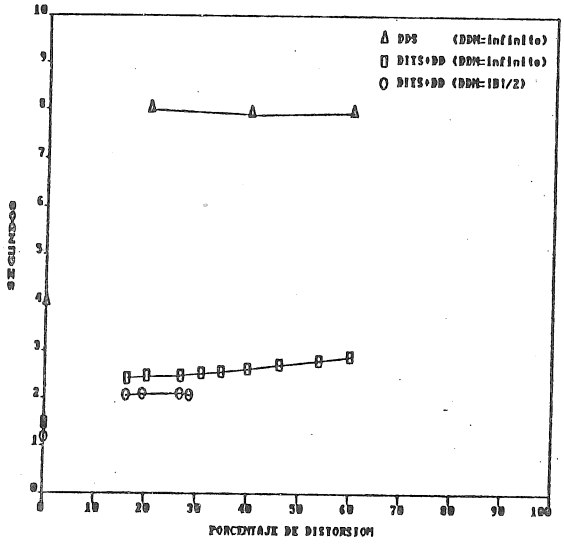


Figura 4 "Gráfica de los tiempos promedios de búsqueda para los esquemas DDS y DITS+DD"

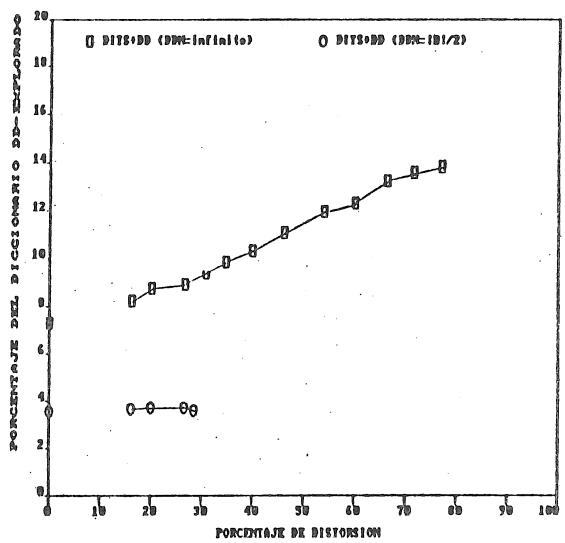


Figura 5 "Gráfica del porcentaje promedio del diccionario implicado en el cálculo de DD en el esquema DITS+DD"

Los tiempos de DDS para búsqueda de palabras correctas, aquellas que se encuentran en el diccionario, son aproximadamente la mitad que para palabras distorsionadas. Ello es debido a que durante la búsqueda lineal, si la palabra no está en el diccionario, éste ha de ser recorrido totalmente, mientras que si la palabra de búsqueda sí que está, se recorre como promedio la mitad del mismo.

La complejidad de cálculo del esquema de búsqueda DDS para palabras distorsionadas es insensible a la distorsión.

La complejidad de cálculo del esquema de búsqueda DITS+DD es del orden del 40% del esquema de búsqueda DDS.

En el esquema de búsqueda DITS+DD (DDM=infinito) aproximadamente el 30% del tiempo de búsqueda se invierte en los cálculos de DD y el 70% en los cálculos de DIT. La reducción del tiempo para este esquema con DDM=|B|/2 se debe a que el número de palabras DD-exploradas se ha reducido a un 40%. Por lo que bajo esta condición el 15% del tiempo se invierte en los cálculos de DD y el 85% restante en los cálculos de DIT.

El ligero aumento del los tiempos promedio en el esquema de búsqueda DITS+DD para palabras distorsionadas, se debe exclusivamente al pequeño incremento del número de palabras del diccionario a las que se le calcula la DD con la palabra de búsqueda.

6.- ESTRUCTURACION DE LAS COMPONENTES QUE INTERVIENEN EN EL CALCULO DE DIT.

Los vectores de frecuencias de los caracteres de las palabras de igual longitud pueden estructurarse permitiendo, conjuntamente con el esquema de búsqueda apropiado, abandonar la exhaustividad de la búsqueda, con lo que sólo se deberá tratar una parte del diccionario para la solución del problema. Se tiene, por tanto, un NODO-RAIZ que discrimina por longitudes de palabras de forma que en la posición i se tiene la dirección de la parte de la estructura que contiene las palabras de longitud i.

6.1 PARTE-ARBOL.

Se puede definir una estructura arbórea cuyos nodos son de la siguiente forma:

α1	10	11	..	1ne

donde:

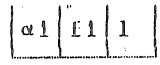
αi: carácter del alfabeto.

1k k=0..ne: un enlace que señala la parte del árbol donde se encuentran las palabras DjεD/djαi=k.

Se llama parte-árbol a esta parte de la estructura.

6.2 PARTE-CADENA.

Cuando por una rama de la parte-árbol ya no se pueda discriminar, se utiliza una lista encadenada formada por nodos de la forma:



αi: carácter del alfabeto que no ha sido utilizado en el camino desde la raiz del árbol.
fi: frecuencia de αi.
l: enlace.

Se llama parte-cadena a esta parte de la estructura.

6.3 CADENA-SIT.

Cuando se acaban los caracteres del vector de frecuencias, tanto en la parte-árbol como en la parte-cadena, se coloca un nodo terminal con αi='/' tal que su enlace apunta a una lista encadenada de sinónimos invariantes trasposicionales, SIT, que son palabras que la estructura considera indiscernibles por tener vectores de frecuencias de caracteres idénticos, aunque éstos tienen una disposición distinta, (AMOR, ROMA, RAMO).

Se llama cadena-SIT a esta parte de la estructura.

En el caso de nuestro diccionario las longitudes de las cadenas-SIT tienen la siguiente distribución de frecuencia:

Long. cad-SIT	1	2	3
Frecuencias	2000	43	1

La longitud promedio es de 1.022

6.4 CRITERIO DE ELECCION DEL CARACTER DISCRIMINANTE EN LA PARTE-ARBOL.

Llega el momento de plantearse que caracteres αi se van eligiendo durante la creación del árbol. Se ha elegido un criterio dinámico; el árbol se va creando según se van insertando palabras. Cuando se hace necesario crear un nodo que separe los caminos de dos palabras no-SIT, se escoge el primer carácter por orden alfabético que establezca una diferencia.

7.- MEDIDAS ESTATICAS.

Creada esta estructura con el diccionario se tomaron las siguientes medidas estáticas:

1.- Longitud de caminos en la parte-árbol.
2.- Longitud de los caminos en la parte-cadena.
3.- Porcentajes de enlaces que están a nulo en cada nivel de la parte-árbol.

En las gráficas de las figuras 6 y 7 se observa una simetría especular debido a que el alfabeto tiene 26 caracteres; si existen t caminos en el árbol de longitud lt entonces también existen t listas encadenadas con longitud 26-lt.

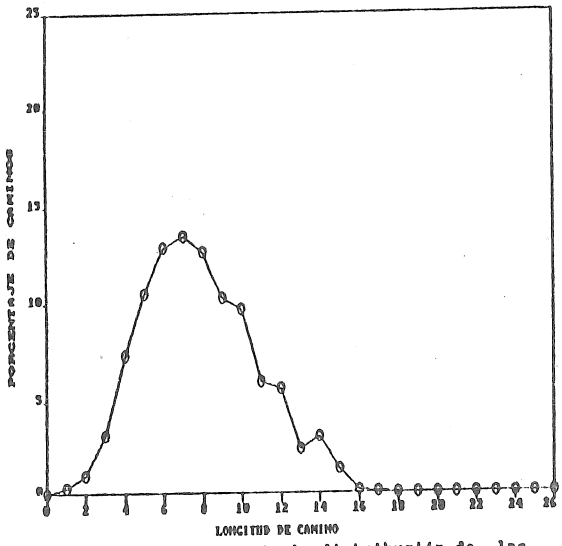


Figura 6 "Gráfica de la distribución de las longitudes de los caminos en la parte-árbol"

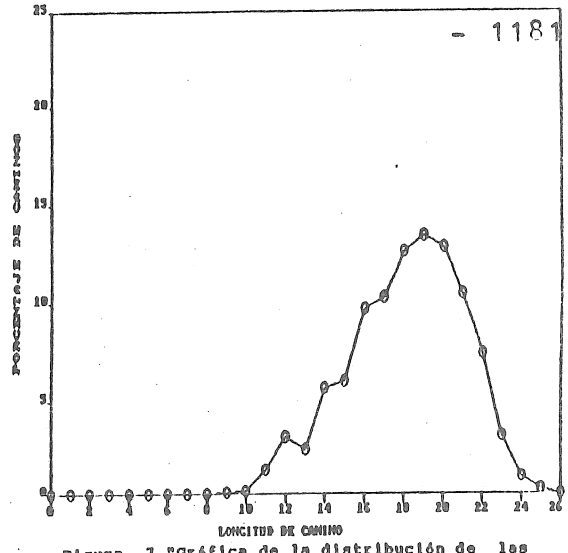


Figura 7 "Gráfica de la distribución de las longitudes de los caminos en la parte-cadena"

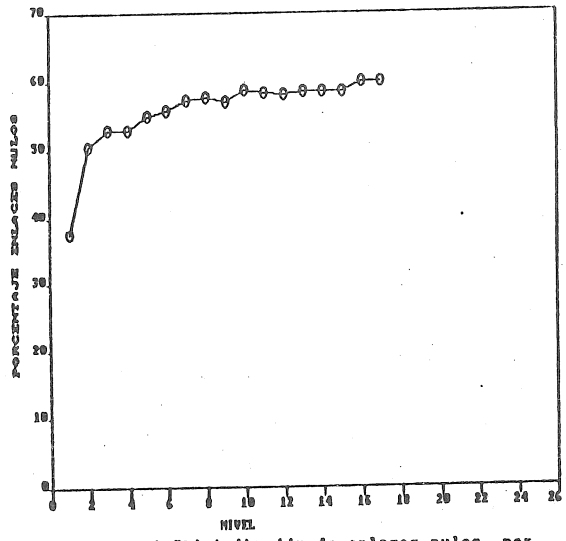


Figura 8 "Distribución de enlaces nulos por niveles en la parte-árbol con ne=4"

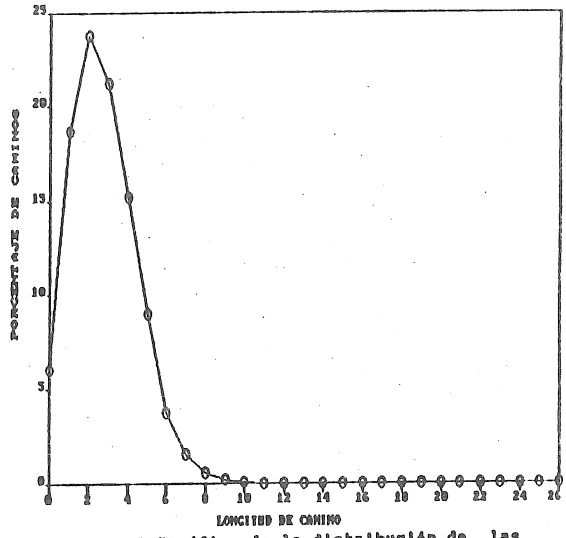


Figura 9 "Gráfica de la distribución de las longitudes de los caminos en la parte-cadena optimizada"

8.- OPTIMIZACION DE LA PARTE-CADENA.

Una mejora inmediata que se puede realizar es eliminar de la lista encadenada (parte-cadena) aquellos nodos cuya fi=0. De esta forma la figura 7 queda como refleja la figura 9.

Se observa, figura 9, que se produce un fuerte decremento de la longitud promedio de la parte-cadena; de 18.115 a 2.944.

9.- OPTIMIZACION DE LA PARTE-ARBOL.

9.1- OPTIMIZACION DEL CRITERIO DE ELECCION DEL CARACTER DISCRIMINANTE EN LA PARTE-ARBOL.

Debido a la naturaleza estática de un diccionario se puede pensar en crear el árbol conociendo a priori que palabras va a contener. De esta forma se puede escoger en todo momento el carácter que más eficazmente realicze su trabajo discriminatorio. ¿Cómo se puede hacer esta elección?. Utilizando un criterio de uniformidad.

Sea N el número de palabras que van a colgar de un nodo. N es independiente del carácter αi que se escoja como discriminante para ese nodo.

Sea αilk el número de palabras que colgarían del enlace k si se escogiese el carácter αi.

$$\sum_{k=0}^{ne} (\alpha ilk) = N \text{ para cualquier } i$$

Se escoge aquel αi tal que:

$$\sum_{k=0}^{ne} (\alpha ilk)^2 \text{ sea mínimo}$$

Esto asegura una distribución uniforme de las palabras a lo largo de cada nodo.

Las figuras 10 y 11 muestran como queda el árbol tras este proceso. Se aprecia un disminución de la altura promedio de la parte-árbol que de 7.885 pasa a 6.032. Además se da mayor agrupamiento en torno a la media como refleja la varianza:

V1=8.519
V2=3.107

Igualmente el número de nodos que tiene la parte-árbol optimizada es un 95.59% del anterior. Además con el primer criterio en la parte-árbol el 56.14% de los enlaces estaba a nulo y ahora el porcentaje es del 55.04%. Se observa igualmente un aumento en el número de listas encadenadas cuya longitud es cero. Todo esto da a entender un mayor aprovechamiento de la estructura.

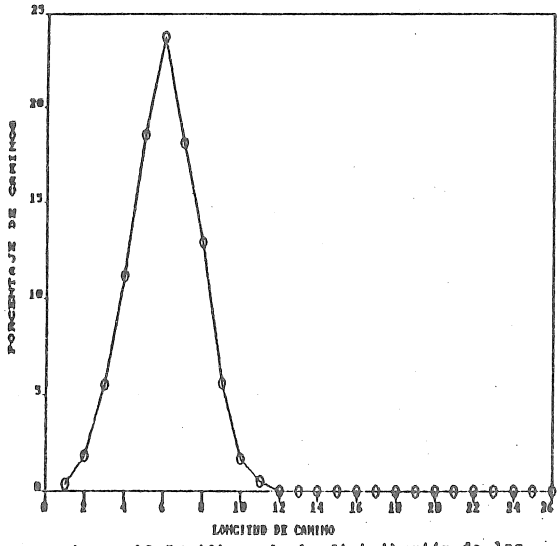


Figura 10 "Gráfica de la distribución de las longitudes de los caminos en la parte-árbol"

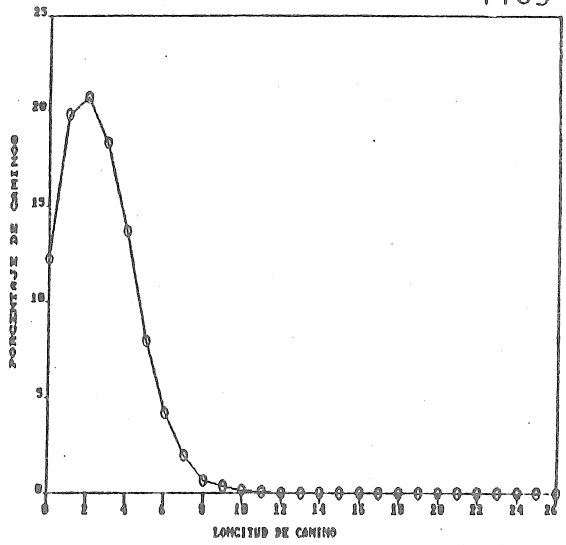


Figura 11 "Gráfica de la distribución de las longitudes de los caminos en la parte-cadena"

En las figuras 12 y 13, se tiene un reflejo de la estructura 'física' de la parte-árbol sin y con optimización.

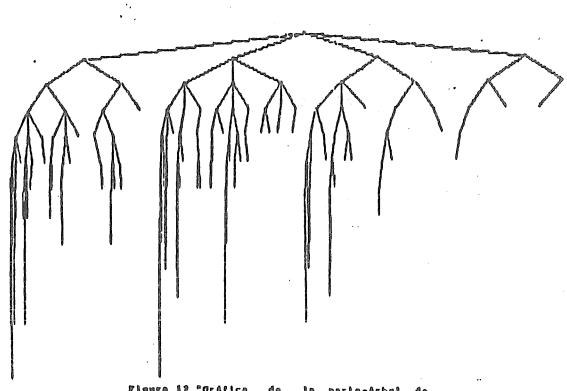


Figura 12 "Gráfica de la parte-árbol de palabras de longitud 7 sin optimización"

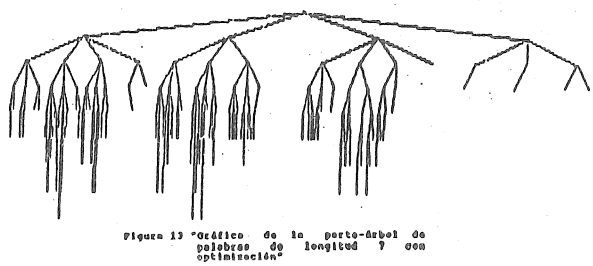


Figura 13 "Gráfica de la parte-árbol de palabras de longitud 7 con optimización"

9.2 ENLACES NULOS. TIPOS DE NODOS.

La siguiente mejora de la estructura viene de observar el elevado porcentaje de enlaces que están a nulos. ¿Podría existir alguna forma de eliminarlos?. Para ello se observó que gran parte de los nodos tenían los enlaces no-nulos agrupados, bordeados de enlaces nulos. Es decir, que de los ne enlaces disponibles en cada nodo, desde 10 hasta un cierto lF y a partir de un lR hasta lne todos los enlaces están a nulos. Por ello se dice que un nodo es de tipo [F,R] si:

{lz=nulo, z∉[F..R], lF y lR≠nulo}

El siguiente cuadro muestra la frecuencia de los distintos tipos de nodos:

Tipo	1	2	3	4
0	73.05%	20.31%	2.95%	0.25%
1		2.34%	0.62%	-----
2			0.18%	0.06%
3			0.06%	-----

La existencia de un nodo de tipo [3,3] es debido a que en el diccionario existe una sóla palabra de longitud dieciocho, por lo que únicamente se necesita un nodo de la parte árbol con el fin de que exista esta parte. Es un caso muy particular que no se ha considerado oportuno incluir su excepcionalidad en el tratamiento de la estructura.

9.3 MODIFICACION DE LOS NODOS.

Se varia ligeramente la definición del nodo de la parte-árbol:

αi	F	R	lF	..	lR

De esta forma se consigue eliminar todos los enlaces nulos excepto los comprendidos entre F y R. Con lo que para este experimento sólo el 0.94% de los enlaces son nulos.

10.- ESQUEMA DE BUSQUEDA DE LAS PALABRAS MAS SIMILARES USANDO LA ESTRUCTURA OPTIMIZADA (DITE+DD).

En el esquema de búsqueda cabe distinguir cuatro partes perfectamente diferenciadas:

1.- Recorrido por el nodo raiz.
2.- Recorrido por la parte-árbol.
3.- Recorrido por la parte-cadena.
4.- Recorrido por la cadena-SIT.

Estas cuatro partes forman una estructura jerárquica de forma que de una parte se pasa a la siguiente por llamada y a la anterior por retorno.

1.-Recorrido por el nodo raiz.-

Se empieza entrando por la posición correspondiente a la longitud de la palabra de búsqueda y luego se van tomando alternativas por proximidad de longitud. Por ejemplo, si $|B|=3$ se explora el nodo raiz mediante la siguiente secuencia:

3,4,2,5,1,6,7,8,9,...

Estas secuencias se tienen almacenadas en una tabla de doble entrada.

ORDEN DE LA SECUENCIA

LONGITUDES	1	2	3	4	..
1	1	2	3	4	..
2	2	3	1	4	..
3	3	4	2	5	..
4	4	5	3	6	..
:	:	:	:	:	

Ahora bien en principio se tiene que DDM=INFINITO y DTM=2*DDM, pero este filtro se va ajustando cuando la búsqueda se ejecuta en los niveles inferiores.

En el cálculo de la DIT uno de los sumandos es la diferencia en valor absoluto de la longitud de la palabra de búsqueda y la longitud de la palabra del diccionario; como en cada subárbol las palabras están agrupadas por longitud, este sumando puede calcularse y entregarse a los niveles inferiores.

Como es lógico suponer, este sumando se utiliza para interrumpir la secuencia de recorrido del nodo raiz, en el caso de que sea mayor que el valor de DDM.

2.- Recorrido de la parte-árbol.-

El vector de frecuencias de B se calcula antes de empezar la búsqueda.

Si se está en un nodo cuyo carácter es α se van recorriendo sus enlaces de un modo parecido a como se hace en el nodo raiz, por alternativas. Pero además hay que tener en cuenta que los nodos están clasificados por tipos. Por ejemplo, si un nodo es de tipo [1,3] y B tiene cuatro veces el caracter α (bα=4) se recorre este nodo según la secuencia 3,2,1; pero si bα=2 la secuencia será 2,3,1.

Cada valor de una secuencia depende de cuatro factores:

1. Límite inferior del tipo de nodo.
2. Límite superior del tipo de nodo.
3. bα.
4. Posición del valor dentro de la secuencia.

Por ello se utiliza una tabla de cuatro entradas.

Cada vez que se accede a un nodo, se distinguen dos casos:

1. Se direcciona por el enlace propio, es decir, por aquél que corresponde según bα.

2. Se direcciona por un enlace alternativo.

Esta distinción es debida a que se va calculando la DIT según se baja por la parte-árbol (DITACUM). En el caso 1. la DITACUM que se tiene calculada no sufre ninguna variación, mientras que en el caso 2. hay que sumar abs(bα-i), siendo i el enlace por el que se direccione.

Si en algún momento DITACUM es superior a DTM no se continua con la toma de alternativas en este nodo, sino que se devuelve el control al nodo que está en el nivel inmediato superior para continuar la búsqueda por las alternativas apropiadas.

3.- Recorrido por la parte-cadena.-

Aquí es un sencillo recorrido secuencial, calculando en cada nodo el término correspondiente de la DIT, si α es el carácter del nodo se le suma a DITACUM el término abs(bα-fi). Si en algún momento se supera la DTM se suspende este recorrido y se le cede control al último nodo visitado de la parte árbol.

4.- Recorrido por la cadena-SIT.-

Una vez agotada la parte-cadena se calcula la DIT en base a DITACUM y a los caracteres de B que no hayan sido utilizados desde la raíz del árbol. Sólo en el caso de que la DIT no supere la DTM se pasa a calcular la DD con las palabras que estén en la cadena-SIT, se compara con DDM, a fin de ver si se ha de añadir la palabra o no al conjunto de palabras más similares y se ajusta, si procede, DDM y DTM.

11 MEDIDAS DINAMICAS.

El tiempo de búsqueda de las palabras más similares a una dada se invierte en:

1. El cálculo de las DD entre la palabra de búsqueda y las del diccionario que superen el filtro de la DIT. Esto supone alrededor del 60% del tiempo de búsqueda.

2. El recorrido por la estructura evaluando las DITs, utiliza alrededor del 40% del tiempo de búsqueda.

Por lo que el esquema DITE+DD consume más tiempo en el cálculo de las DD que en los de DIT, contrariamente a lo que ocurría en el esquema DITS+DD.

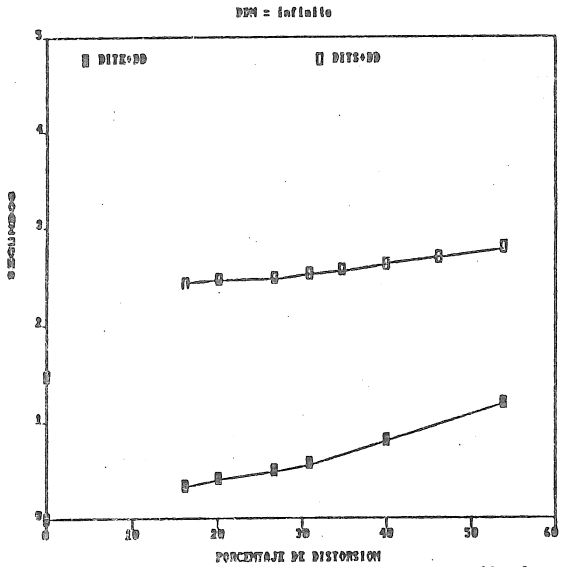


Figura 14 "Gráfica de los tiempos promedio de búsqueda para los esquemas DITS+DD y DITE+DD "

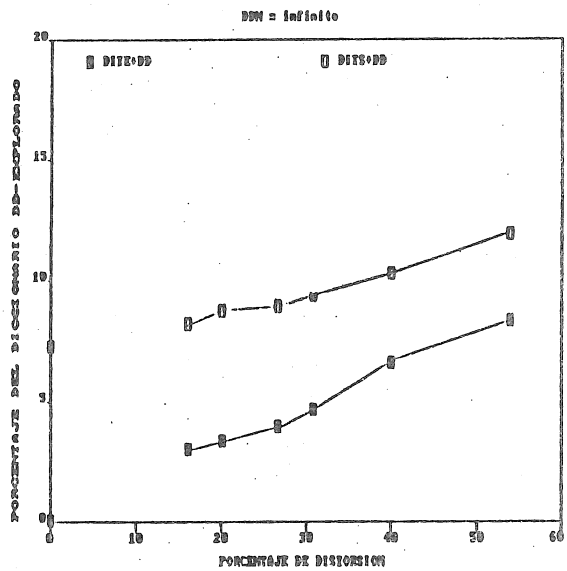


Figura 15 "Gráfica del porcentaje promedio del diccionario implicado en el cálculo de DD en los esquemas DITS+DD y DITE+DD"

Respecto al esquema de búsqueda secuencial DITS+DD (DDM=infinito) para la búsqueda de palabras distorsionadas se ha reducido el número de palabras del diccionario que han de intervenir en el cálculo de la DD, equivalente a un 20% del ahorro de tiempo por palabra de búsqueda. De lo que se deduce que alrededor del 80% de la optimización en tiempo obtenida proviene de la menor sobrecarga del cálculo de la DIT al estructurar sus componentes.

12.- MODIFICACION ULTERIOR AL ESQUEMA ESTRUCTURADO

Se introducirá una modificación al esquema de búsqueda en la estructura, proponiendo un nuevo esquema que en adelante se denominará DITEM+DD. Dicha modificación consiste en cambiar el sentido aplicado hasta ahora al juego de inicializaciones de DDM, de forma que se utilizará una afirmación mas restrictiva, garantizando que el canal no producirá una distorsión mayor que un cierto porcentaje. Por lo tanto, aquellas palabras del diccionario con respecto a las cuales la palabra de búsqueda tenga una distorsión superior a la fijada, serán eliminadas automáticamente por el nuevo esquema. Para ello, se tendrán en cuenta las características de cada rama del nodo raíz de la estructura, con respecto a la longitud de las palabras del diccionario, de forma que se inicializa el DDM al mínimo entre la distancia direccional actual y el producto de la distorsión permitida π por la longitud de las palabras a explorar.

En las figuras 16 y 17 se muestran las gráficas del tiempo promedio de búsqueda y del porcentaje promedio DD-explorado del diccionario, para los esquemas DITE+DD, en sus inicializaciones de DDM a infinito y $\lfloor B/2 \rfloor$ y el esquema DITEM+DD con $\pi=1/3$ (33.33%). En ellas aparece registrado un fuerte descenso en el tiempo promedio de búqueda para la estructura DITEM+DD, debido fundamentalmete a la disminución observada en el porcentaje promedio DD-explorado del diccionario.

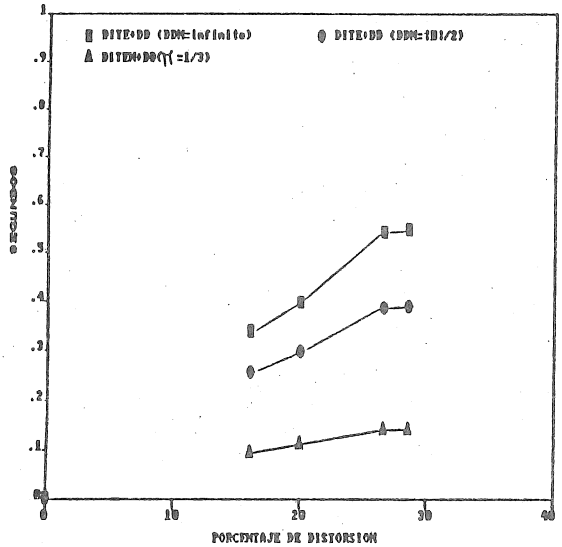


Figura 16 "Gráfica de los tiempos promedio de búsqueda para los esquemas DITE+DD y DITEM+DD"

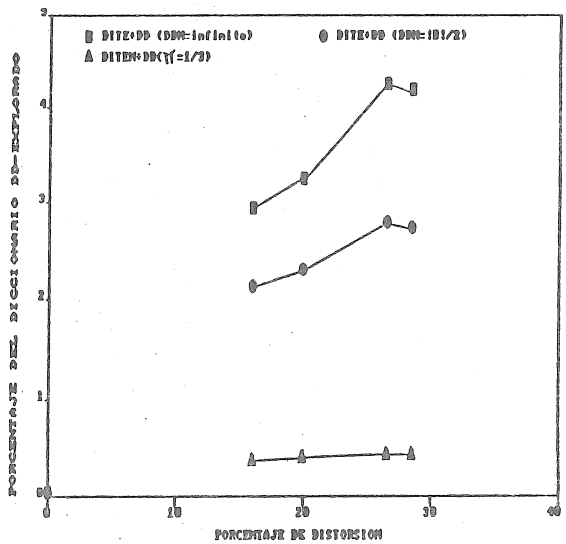


Figura 17 "Gráfica del porcentaje promedio del diccionario implicado en el cálculo de DD en los esquemas DITE+DD y DITEM+DD"

13 CONCLUSIONES.

La evolución del tema tratado aquí, parte del tratamiento más simple posible: un recorrido secuencial, exhaustivo en el caso de una palabra de búsqueda incorrecta, con calculo de la distancia direccional (DDS). Este tratamiento proporciona un porcentaje promedio de aciertos razonablemente alto hasta distorsiones del 30%, mientras que el porcentaje promedio de respuesta única no es muy alto, incluso para distorsiones bajas, y el promedio de multiplicidad de respuesta crece rápidamente a partir del 20% de distorsión.

La introducción el esquema DITS+DD, con el fin de mejorar el tiempo promedio de búsqueda utilizando un esquema más sofisticado, produce, en efecto, una disminución del mismo de alrededor del 40%, aunque una pequeña parte de la disminución se debe al aumento del coste de almacenamiento, al guardar los vectores de frecuencias. A diferencia del esquema DDS, en DITS+DD el tiempo promedio de búsqueda para palabras incorrectas se muestra sensible a la distorsión. Asimismo, la inicialización de la distancia direccional máxima a la mitad de la longitud de la palabra de búsqueda, produce una fuerte disminución del porcentaje de diccionario DD-explorado, arrastrando consigo al tiempo promedio de búsqueda.

La estructuración inicial de los datos en el esquema DITE+DD lleva asociada una serie de medidas estáticas que sufren sucesivas mejoras con las tres optimizaciones introducidas a la estructura. Hay que destacar la efectividad de la eliminación de un altísimo porcentaje de encadenamientos nulos y la progresiva disminución de la longitud promedio de los dos tipos de caminos de la estructura.

En cuanto a las medidas dinámicas tomadas al esquema DITE+DD hay que destacar una fuerte bajada del tiempo promedio de búsqueda, con respecto al del esquema DITS+DD, así como la inversión de la importancia relativa de sus componentes: al contrario que en el esquema anterior, en el DITE+DD tiene más peso el tiempo invertido en el cálculo de las distancias direccionales. Otro factor a destacar es la disminución del porcentaje promedio del diccionario que es DD-explorado, con respecto al esquema DITS+DD. Este factor equivale a una pequeña parte del total de la disminución del tiempo de búsqueda, por lo que ha de entenderse que el peso de esta disminución recae sobre la menor sobrecarga en el cálculo de las distancias invariantes trasposicionales, a causa de la estructuración. De nuevo, tanto el tiempo promedio de búsqueda, como el porcentaje DD-explorado mejoran al inicializar DDM a la mitad de la longitud de la palabra de búsqueda.

La modificación introducida en el esquema DITEM+DD conlleva una fuerte disminución de los tiempos de búsqueda y una bajada del porcentaje DD-explorado con respecto al esquema DITE+DD. Por último, la respuesta al uso de diferentes distorsiones máximas sigue las líneas trazadas en los anteriores esquemas con el uso de diferentes inicializaciones de DDM, con la salvedad de que en este caso, habrán de ser rechazadas las posibles palabras propuestas como solución por tener una distancia direccional baja, que superen la distorsión máxima permitida.

BIBLIOGRAFIA

{1} AHO, A.V.; HOPCROFT, J.E.; ULLMAN, J.D.: Data Structures and Algorithms. Addison-Wesley Publishing Company. 1983

{2} ALBERGA, C.N.: String Similarity and Misspellings. CACM 1967, 10 (5),302 313.

{3} MORGAN, H.L.: Spelling Correction in System Programs. CACM 1970, 13 (2), 90-94

{4} SZANSER, A.J.M.: Automatic Error Correction of Natural Text. Part I Computer Science N. 46 National Physics Laboratory, 1973.

{5} TREMBLAY, J.P.; SORENSON, P.G.: An Introduction to Data Structures with Applications. McGraw-Hill Book Company. 1984.

{6} WAGNER, R.A.; FISCHER, M.J.: The String-to-String Correction Problem. JACM 1974, 21 (1), 168-173.