

**UNIVERSIDAD DE
LAS PALMAS DE GRAN
CANARIA
E. T. S. I. TELECOMUNICACIÓN**



TESIS DOCTORAL

**Estructura y esquemas de
búsqueda por similitud de
cadenas de caracteres.**

**Una aplicación para peticiones complejas
de localización de palabras en archivos
documentales**

Autora: D^a. Margarita Díaz Roca

Director: Dr. D. Octavio Santana Suárez

Departamento de Electrónica y
Telecomunicación
Mayo 1993

UNIVERSIDAD DE LAS PALMAS DE GRAN CANARIA

DOCTORADO EN INGENIERÍA DE TELECOMUNICACIÓN

DEPARTAMENTO DE ELECTRÓNICA Y TELECOMUNICACIÓN
PROGRAMA DE INGENIERÍA ELECTRÓNICA

**ESTRUCTURA Y ESQUEMAS DE
BÚSQUEDA POR SIMILITUD DE CADENAS
DE CARACTERES.**

**UNA APLICACIÓN PARA PETICIONES COMPLEJAS DE
LOCALIZACIÓN DE PALABRAS EN ARCHIVOS
DOCUMENTALES**

*Tesis Doctoral presentada por D^a. Margarita Díaz Roca
Dirigida por el Dr. D. Octavio Santana Suárez*

El Director,

La Doctoranda,

Las Palmas de Gran Canaria. Mayo, 1993

RESUMEN

Este trabajo trata aspectos teóricos y experimentales en torno al problema de la búsqueda de las cadenas más similares a una dada. El concepto de similitud es en el sentido de la distancia de Levenshtein, DL . El objetivo que se persigue es la optimización de los recursos de tiempo y espacio de los esquemas de búsqueda y de la estructura de datos que los soporta.

Se define una nueva distancia que se ha denominado distancia invariante trasposicional, DIT , debido al hecho de que su valor no depende de las operaciones de trasposición a que pueda ser sometida una cadena. Si bien DIT no puede usarse por sí sola para la determinación de las cadenas más similares, su importancia deviene de la circunstancia de que su valor entre dos cadenas es siempre inferior o igual a la DL entre estas dos mismas cadenas, siendo su coste computacional sensiblemente inferior; lo cual puede ser aplicado para la construcción de un filtro adaptivo DIT/DL que tenga por misión reducir el número de cadenas de la base de datos a las que se les calcula la DL con la cadena de búsqueda.

Se diseña una estructura, $S-D$, al objeto de compartir las componentes de DIT y no tener que calcular completamente la DIT de la cadena de búsqueda a todas y cada una de las cadenas del diccionario. El esquema de búsqueda de las cadenas más similares que se apoya en esta estructura, recorriéndola a través de las componentes de DIT , y que usa este valor como criterio de poda se denomina esquema decreciente. Se estudian nuevas estrategias para un esquema de búsqueda creciente, donde el radio de búsqueda, en oposición a la evolución clásica decreciente, sigue una línea de modificación creciente. Asimismo, se propone un esquema decreciente con radio ascendente tal que en función del incremento del radio de búsqueda define una familia de esquemas intermedios que conectan a los esquemas creciente y decreciente.

Prolongando la línea de optimización de las realizaciones de los esquemas de búsqueda decreciente y creciente, se define un nuevo umbral DS , cuyo valor se encuentra entre DIT y DL ; y debido a que tiene un menor costo que DL es útil para descartar un número de cadenas a las que no es necesario evaluar DL . También se introduce un refinamiento en la poda del índice que acorta su recorrido.

Si el tamaño de la estructura es tal que no es posible ubicarla en

memoria interna, se plantea el problema de su paginación con el fin de minimizar el número de accesos a disco. Un primer intento para resolver el problema consiste en llenar las páginas con los nodos al recorrer la estructura en preorden o en postorden. También se propone una nueva forma de paginar la estructura de tal modo que se respete el siguiente principio: durante la búsqueda, una vez que se abandone una página no se vuelve a acceder a ella por otro camino.

Finalmente se concreta en una aplicación a la localización de palabras en texto libre, que constituye una parte fundamental de un problema práctico de gran importancia: la organización y utilización de información procedente de fuentes heterogéneas. Existen dos aproximaciones a este problema, efectuar la búsqueda directamente sobre el documento sin ningún tipo de preparación previa o someter los escritos a un tratamiento anterior y generar un índice que haga viable los accesos posteriores al ejemplar. Siguiendo este último enfoque, se utiliza como índice la estructura *S-D* construida a partir del conjunto de las palabras diferentes que se obtienen del documento al descartar las cadenas consideradas *vacías*. Sobre la estructura así construida se permiten los siguientes tipos de búsquedas: exacta, más similares, con operadores booleanos, máscaras, truncamientos, cercanía, antecendencia, párrafos, sentencias, frases y complejas.

Se construye asimismo un analizador sintáctico que cumple un doble cometido, ya que ha de identificar cada tipo de petición –diferenciando las componentes y los conectores lógicos en el caso de las complejas– y, a la vez, determinar la correctitud sintáctica. Como complemento al analizador se realiza un optimizador para solucionar las búsquedas complejas en el menor tiempo posible; se pretende calcular una petición equivalente a la solicitada por el usuario de forma que se resuelva más rápidamente, minimizando así el tiempo de respuesta del sistema.

A Paco, Aída y Teba

AGRADECIMIENTOS

Quiero expresar mi profundo agradecimiento al Profesor D. Octavio Santana Suárez, Director del Grupo de Investigación en Estructuras de Datos en el cual se ha desarrollado mi trabajo, por haber desempeñado un papel esencial en mi formación científica y profesional. Sin sus reflexiones, consejos y dirección no hubiera sido posible realizar esta Tesis.

A todos mis compañeros del Grupo de Investigación en Estructuras de Datos, mi agradecimiento especial por haberme animado siempre. Gracias por su colaboración, ayuda incondicional y disponibilidad con la que siempre he podido contar.

También quiero agradecer al Departamento de Electrónica y Telecomunicación y a la Escuela Técnica Superior de Ingenieros de Telecomunicación de la Universidad de Las Palmas de Gran Canaria, y en particular a su Director, el Profesor D. Antonio Núñez Ordóñez, el interés y el apoyo mostrado para llevar a buen término este trabajo.

ÍNDICE

Introducción	1
CAPÍTULO 1: Distancias entre Cadenas	9
1.1 Distancia de Levenshtein, DL.	10
1.2 Cálculo de Wagner y Fischer de la Distancia de Levenshtein.	11
1.3 Cálculo Optimizado de la Distancia de Levenshtein.	13
1.4 Distancia Invariante Trasposicional, DIT.	19
CAPÍTULO 2: Estructura de Santana y Díaz, S-D, y Esquemas de Búsqueda de las Más Similares	25
2.1 Estructura de Santana y Díaz, S-D.	26
2.1.1 Criterios de Selección del Carácter Discriminante. ..	29
2.1.2 Relación Ocupacional.	31
2.2 Búsqueda de las Cadenas Más Similares.	33
2.2.1 Esquema de Búsqueda Decreciente.	34
2.2.2 Familia de Esquemas.	38
2.2.3 Esquema de Búsqueda Creciente.	40
2.2.4 Resultados Experimentales.	44
CAPÍTULO 3. Optimización de los Esquemas de Búsqueda Decreciente y Creciente	54
3.1. Podas en el Índice: PA y PP.	54
3.1.1. Resultados Experimentales.	56

	ix
3.2. Distancia de Santana, DS.	58
3.2.1. Resultados Experimentales.	68
CAPÍTULO 4. Paginación de la Estructura S-D	72
4.1. Paginación Según el Recorrido.	73
4.1.1. Resultados Experimentales.	75
4.2. Paginación Atendiendo al Esquema de Búsqueda Decreciente.	82
4.2.1. Resultados Experimentales.	85
CAPÍTULO 5. Aplicación de la Estructura S-D a la Recuperación de Información en Archivos Documentales	88
5.1. Tipos de Búsquedas en Archivos Documentales.	91
5.2. Búsqueda Exacta.	92
5.3. Búsqueda de las Palabras Más Similares.	93
5.4. Búsqueda con Máscaras.	94
5.5. Búsquedas con Truncamientos.	97
5.6. Búsquedas con Operadores Booleanos.	101
5.6.1. Búsqueda Disyuntiva.	101
5.6.2. Búsqueda Conjuntiva.	104
5.6.3. Búsqueda con Operador y_no.	106
5.7. Búsquedas de Cercanía y Antecedencia.	107
5.7.1. Cercanía.	107
5.7.2. Antecedencia.	108
5.8. Búsquedas en Párrafos y Sentencias.	108
5.8.1. Párrafos.	109

5.8.2. Sentencias.	109
5.9. Búsqueda de Frases.	109
5.10. Búsqueda Compleja.	110
5.10.1. Búsqueda en la que intervienen Peticiones Anteriores.	111
5.11. Analizador Sintáctico de la Petición.	111
5.12. Optimizador de la Petición.	112
5.13. Resultados Experimentales. Ubicación del Índice en Memoria Interna.	114
5.14. Resultados Experimentales. Índice Paginado.	119
CAPÍTULO 6. Conclusiones y Principales Aportaciones .	131
Apéndices	
Apéndice 1. Lista de Palabras Vacías	135
Apéndice 2. Abreviaturas y Siglas	142
Referencias	144