

# Informe simulador de búsqueda

Miguel Maripillan<sup>†</sup>, Diego Sanhueza<sup>†</sup>, Miguel Loaiza<sup>†</sup> y Felipe Cárcamo<sup>†</sup>

<sup>†</sup> Universidad de Magallanes

This manuscript was compile on 19 de noviembre de 2024

## Resumen

Este trabajo se enfoca en construir un simulador de un sistema de recuperación de información usando estructuras de datos como; *grafos*, *listas enlazadas*, *tablas hash*, entre otras estructuras de datos que nos serán de utilidad. Se utilizarán principalmente dos técnicas para manejar datos, *índice invertido* y *pagerank*. Este trabajo fue programado usando el lenguaje C.

**Keywords:** PageRank, Simulador, Índice invertido, búsqueda

## ■ Índice

1	Introducción	2
2	Entendiendo el problema	2
2.1	Estructuras de datos	2
2.2	Algoritmos	2
2.3	Resultado del simulador	2
3	Definiendo estructuras y funciones	2
3.1	Prototipo de las funciones principales	2
3.2	Estructuras	3
4	Explicación de las funciones principales	4
4.1	PageRank	4
4.2	Construcción del índice invertido	5
4.3	Conexión de las aristas con los vértices	6
4.4	Creación de archivos	7
5	Hilo de ejecución y salida del programa	10
5.1	Entrada del programa	10
	src2 Creación de Archivos • src Simulador de búsqueda	
5.2	Salida del programa	12
6	Dificultades enfrentadas	13
6.1	Hasheo y tokenización	13
6.2	Generación de la Imagen EPS	13
6.3	Separación de archivos	13
7	Conclusión	13

## 1. Introducción

El proyecto se enfoca en crear un índice invertido para hacer búsquedas rápidas de palabras clave y en construir un grafo dirigido que modela las relaciones entre documentos, usando el algoritmo PageRank para calcular su relevancia. Todo esto busca ayudarnos a entender mejor cómo se organiza y clasifica la información, usando estructuras de datos prácticas y algoritmos avanzados de búsqueda, pero de una manera accesible y directa.

## 2. Entendiendo el problema

Para poder implementar el simulador, se nos piden los siguientes requerimientos;

### 2.1. Estructuras de datos

Usar listas enlazadas o tablas hash para el índice invertido, representar el grafo de enlaces entre documentos mediante listas de adyacencias.

### 2.2. Algoritmos

Implementar el algoritmo PageRank en iteración usando un factor de amortiguación (*damping factor*), tener un algoritmo de búsqueda eficiente en el índice invertido.

### 2.3. Resultado del simulador

Al finalizar el programa se debe mostrar una lista de documentos ordenados por relevancia dado una palabra o frase.

## 3. Definiendo estructuras y funciones

Para poder implementar las funciones requeridas, primero se define un archivo .h con los prototipos de las funciones y las estructuras necesarias.

### 3.1. Prototipo de las funciones principales

```

1
2 1
3 2 /* Funciones Grafos */
4 3 void initialize_graph(Graph *);
5 4 void add_edge(Graph *, int, int);
6 5 void build_graph(Graph *);
7 6 void release_graph(Graph *);
8 7 int count_output_links(Graph *, int);
9 8 int count_input_links(Graph *, int);
10 9 int get_doc_id(Graph *, char *);
11 10 bool is_doc_name(char *);
12 11 void show_graph(Graph *);
13 12 void generate_eps(const Graph *, const double *, const char *);
14 13
15 14 /* Funciones Indice invertido */
16 15 InvertedIndex *create_new_node(char *);
17 16 void add_document(InvertedIndex **, int, char *);
18 17 void tokenize_text(char *, int, InvertedIndex **);
19 18 void print_inverted_index(InvertedIndex *);
20 19 unsigned int hash_function(char *);
21 20 Node *search_word(InvertedIndex **, char *);
22 21 void print_search_word(InvertedIndex **, char *);
23 22 void build_index(Graph *, InvertedIndex **);
24 23 void release_inverted_index(InvertedIndex **);
25 24
26 25 /* Funciones PageRank */
27 26 void initialize_pagerank(double *, int);
28 27 void calculate_pagerank(Graph *, double *);
29 28 void display_pagerank(Graph *, double *);
30 29
31 30 /* Funciones para generacion de archivos */
32 31 void generate_text_files(int, int);
33 32 void generate_random_text(FILE *, const char *, int, int, int, int *);

```

**Código 1.** Prototipo de las funciones.

Estas funciones nos ayudarán a ejecutar las distintas tareas que el simulador debe realizar, mas adelante serán explicadas las funciones principales.

### 3.2. Estructuras

```

1 // Estructura Graph que representa el grafo de documentos con enlaces entrantes y salientes.
2 typedef struct Graph
3 {
4     Node *output_adjacent_list[MAX_DOCS]; // Lista de enlaces salientes.
5     Node *input_adjacent_list[MAX_DOCS]; // Lista de enlaces entrantes.
6     DocumentMapping mapping_docs[MAX_DOCS]; // Mapeo de documentos.
7     int total_docs; // Total de documentos en el grafo.
8 } Graph;

```

**Código 2.** Estructura del grafo.

**Graph:** Esta estructura representa el grafo de documentos y los enlaces entre ellos, implementado mediante listas de adyacencia.

**Atributos:**

- Node \*output\_adjacent\_list[MAX\_DOCS]: Array de listas enlazadas para enlaces salientes desde cada documento.
- Node \*input\_adjacent\_list[MAX\_DOCS]: Array de listas enlazadas para enlaces entrantes a cada documento.
- DocumentMapping mapping\_docs[MAX\_DOCS]: Mapeo de nombres de documentos a identificadores únicos.
- int total\_docs: Número total de documentos representados en el grafo.

```

1 // Estructura InvertedIndex para el índice invertido, almacenando palabras y los documentos donde
2 // aparecen.
3 typedef struct InvertedIndex
4 {
5     char word[MAX_WORD_SIZE]; // Palabra almacenada.
6     Node *docs_list; // Lista de documentos donde aparece la palabra.
7     struct InvertedIndex *next; // Siguiete nodo en la lista enlazada del índice.
8 } InvertedIndex;

```

**Código 3.** Estructura del índice invertido.

**InvertedIndex:** Esta estructura representa un nodo en una lista enlazada para el índice invertido, el cual almacena palabras y los documentos en los que aparecen.

**Atributos:**

- char word[MAX\_WORD\_SIZE]: Palabra indexada.
- Node \*docs\_list: Lista enlazada de documentos donde aparece la palabra.
- struct InvertedIndex \*next: Puntero al siguiente nodo en la lista enlazada del índice.

```

1 // Estructura Node para la lista enlazada de enlaces en el grafo.
2 typedef struct Node
3 {
4     int doc_id; // Identificador del documento enlazado.
5     struct Node *next; // Puntero al siguiente nodo.
6 } Node;

```

**Código 4.** Estructura del nodo.

**Node:** Esta estructura representa un nodo en una lista enlazada que almacena enlaces entre documentos en el grafo. Se utiliza tanto en la lista de adyacencia de entrada como en la de salida.

**Atributos:**

- int doc\_id: Identificador del documento enlazado, representando un vértice en el grafo.
- struct Node \*next: Puntero al siguiente nodo en la lista, permitiendo crear una lista enlazada de enlaces.

```

63 1 // Estructura para mapear el nombre de documentos a sus identificadores unicos.
64 2 typedef struct DocumentMapping
65 3 {
66 4     char name[MAX_NAME_WEB]; // Nombre del documento.
67 5     int doc_id; // Identificador unico del documento.
68 6 } DocumentMapping;

```

Código 5. Estructura del mapeo de documentos.

**DocumentMapping:** Mapea los nombres de los documentos a identificadores únicos, lo cual es útil para gestionar documentos en el grafo sin confusión.

**Atributos:**

- `char name[MAX_NAME_WEB]`: Nombre del documento (archivo) en el grafo.
- `int doc_id`: Identificador único asociado al documento para fácil referencia.

## 4. Explicación de las funciones principales

### 4.1. PageRank

Este módulo implementa el algoritmo de PageRank para calcular la importancia relativa de cada documento en un grafo web, utilizando enlaces como referencia para el cálculo de la influencia entre documentos. La implementación consta de tres funciones principales: `initialize_pagerank()`, `calculate_pagerank()`, y `display_pagerank()`.

- **`initialize_pagerank()`:** Esta función se encarga de inicializar el vector de PageRank. Cada documento comienza con un valor uniforme de  $1 / \text{total\_docs}$ , de modo que todos los documentos tengan el mismo PageRank inicial. Esto es importante para asegurar que el cálculo sea equitativo desde el principio.
- **`calculate_pagerank()`:** Esta función calcula iterativamente el PageRank para cada documento en el grafo hasta que el cálculo converja o hasta alcanzar el número máximo de iteraciones, determinado por la constante `MAX_ITERATIONS`.

El proceso de cálculo se realiza de la siguiente manera:

1. Se inicializa el vector `new_pagerank` para almacenar los valores de PageRank temporales en cada iteración.
2. Para cada documento, se calcula la suma de los valores de PageRank de los documentos que tienen enlaces hacia él (inbound links). Este cálculo se realiza dividiendo el valor de PageRank del documento que apunta por su cantidad de enlaces salientes, de modo que cada enlace contribuye proporcionalmente a los documentos de destino.
3. Se utiliza un factor de amortiguación (`DAMPING_FACTOR`) para calcular el nuevo valor de PageRank. Este factor permite ajustar la probabilidad de que un usuario siga navegando a través de los enlaces del documento. La fórmula para actualizar el PageRank es:

$$\text{nuevo\_pagerank}[i] = \frac{(1 - \text{DAMPING\_FACTOR})}{\text{graph} \rightarrow \text{total\_docs}} + \text{DAMPING\_FACTOR} \times \text{inbound\_rank\_sum}$$

4. Después de cada actualización de PageRank, se verifica la convergencia comparando el cambio en el PageRank de cada documento con un umbral, `CONVERGENCE_THRESHOLD`. Si todos los cambios están por debajo del umbral, el proceso de iteración se detiene y el cálculo se considera convergido.
5. Al finalizar cada iteración, el vector original `pagerank` se actualiza con los valores calculados en `new_pagerank`.

Si el cálculo converge antes de alcanzar el número máximo de iteraciones, se imprime un mensaje indicando el número de iteraciones necesarias para la convergencia.

- **`display_pagerank()`:** Esta función muestra los resultados del PageRank para cada documento. Imprime en la consola el ID de cada documento junto con su valor de PageRank calculado, permitiendo así evaluar la importancia relativa de cada documento en el conjunto.

```

1 // Calcula el PageRank para cada documento en el grafo.
2 void calculate_pagerank(Graph *graph, double pagerank[])
3 {
4     double new_pagerank[MAX_DOCS]; // Vector temporal para almacenar los nuevos
5     // valores de PageRank en cada iteración.
6     initialize_pagerank(pagerank, graph->total_docs); // Inicializa el PageRank.
7
8     // Itera para calcular el PageRank hasta que converja o se alcance el máximo de iteraciones.
9     for (int iter = 0; iter < MAX_ITERATIONS; iter++)
10     {
11         bool converged = true;
12
13         // Calcula el nuevo PageRank para cada documento.
14         for (int i = 0; i < graph->total_docs; i++)
15         {
16             double inbound_rank_sum = 0.0;
17
18             // Recorre los documentos que enlazan al documento 'i'.
19             Node *inbound = graph->input_adjacent_list[i];
20             while (inbound != NULL)
21             {
22                 int inbound_doc_id = inbound->doc_id; // ID del documento de
23                 // entrada.
24                 int out_links = count_output_links(graph, inbound_doc_id); // Número de enlaces
25                 // salientes desde el documento.
26
27                 // Si el documento tiene enlaces salientes, acumula el PageRank proporcional.
28                 if (out_links > 0)
29                     inbound_rank_sum += pagerank[inbound_doc_id] / out_links;
30
31                 inbound = inbound->next;
32             }
33
34             // Calcula el nuevo valor de PageRank utilizando el factor de amortiguación.
35             new_pagerank[i] = (1 - DAMPING_FACTOR) / graph->total_docs + DAMPING_FACTOR *
36             inbound_rank_sum;
37
38             // Verifica si el cambio entre el nuevo y el viejo PageRank supera el umbral de convergencia.
39             if (fabs(new_pagerank[i] - pagerank[i]) > CONVERGENCE_THRESHOLD)
40                 converged = false;
41         }
42
43         // Actualiza el vector de PageRank con los nuevos valores calculados.
44         for (int i = 0; i < graph->total_docs; i++)
45             pagerank[i] = new_pagerank[i];
46
47         // Si todos los valores de PageRank han convergido, finaliza el cálculo.
48         if (converged)
49         {
50             printf("PageRank convergido en %d iteraciones.\n", iter + 1);
51             break;
52         }
53     }
54 }

```

Código 6. Extracto del código del PageRank.

Se obtuvo información de Wikipedia: [1], igual se visitó GeeksforGeeks [3] y se revisó la página Memgraph [4] para documentarse acerca del pagerank.

## 4.2. Construcción del índice invertido

El índice invertido es una estructura de datos que permite asociar cada palabra con los documentos en los que aparece, facilitando búsquedas rápidas en un conjunto de textos. A continuación, se describe cómo el programa construye y gestiona este índice invertido paso a paso:

1. **Crear un nodo para una nueva palabra:** La función `create_new_node` es responsable de crear un nuevo nodo en el índice invertido para cada palabra que no se ha encontrado previamente. Este nodo almacena la palabra y una lista de documentos donde aparece. Cuando se detecta una palabra nueva, esta función crea el espacio necesario en memoria para el nodo, inicializa la lista de documentos como vacía, y prepara el nodo para ser agregado al índice.

2. **Agregar documentos al índice:** Cada vez que se encuentra una palabra en un documento, la función `add_document` revisa si ya existe un nodo para esa palabra en el índice. Si la palabra ya tiene un nodo, simplemente agrega el identificador del documento (o `doc_id`) a la lista de documentos de esa palabra. Si la palabra es nueva, crea un nodo y lo añade a la tabla hash del índice, permitiendo que en el futuro se pueda ubicar la palabra y sus documentos de forma rápida.

**3. Ignorar palabras irrelevantes:** Para evitar sobrecargar el índice con palabras comunes y sin valor de búsqueda, la función `is_stopword` identifica palabras irrelevantes (*stopwords*) como “a” o “e”. Estas palabras son ignoradas para hacer el índice más eficiente y relevante, asegurando que solo se incluyan términos útiles para las búsquedas.

**4. Procesamiento de texto para indexación:** La función `tokenize_text` prepara el texto del documento para su indexación. Primero, convierte todo el texto a minúsculas y elimina signos de puntuación, lo que permite que palabras como “Hola” y “hola” sean tratadas como iguales. Luego, el texto se divide en palabras o “tokens”, y cada palabra que no sea una *stopword* se añade al índice con la función `add_document`.

**5. Cálculo de posiciones de palabras:** Para ubicar cada palabra de manera rápida, se utiliza una función de *hash* (`hash_function`) que convierte la palabra en un número único. Este número sirve como índice para almacenar y recuperar el nodo de cada palabra en la tabla hash, optimizando la eficiencia de las búsquedas.

**6. Buscar palabras en el índice:** La función `search_word` permite buscar documentos asociados a una palabra en el índice. Dada una palabra, se calcula su posición en la tabla usando la función de hash y se revisa si existe en el índice. Si la palabra está presente, se devuelve la lista de documentos donde aparece; si no, se retorna un valor nulo.

**7. Liberación de memoria:** Finalmente, la función `release_inverted_index` libera toda la memoria utilizada por el índice invertido cuando ya no se necesita. Esta función recorre cada palabra y su lista de documentos, asegurándose de liberar cada nodo y evitar fugas de memoria en el programa.

De esta manera, el índice invertido permite construir un mapa eficiente de palabras a documentos, optimizando así las búsquedas de palabras en grandes conjuntos de texto. Utiliza técnicas de hash y listas enlazadas para asociar cada palabra con los documentos donde aparece, todo en un formato de búsqueda rápida y organizado.

```

128 1 // Agrega un documento al índice invertido, asociándolo con una palabra específica.
129 2 void add_document(InvertedIndex **hash_table, int doc_id, char *word)
130 3 {
131 4     // Calcula el índice en la tabla hash para la palabra.
132 5     unsigned int index = hash_function(word);
133 6     InvertedIndex *current = hash_table[index];
134 7
135 8     // Busca la palabra en el índice.
136 9     while (current != NULL)
137 10     {
138 11         // Si la comparación se cumple, la palabra ya está en el índice
139 12         if (strcmp(current->word, word) == 0)
140 13         {
141 14             Node *new_doc = (Node *)malloc(sizeof(Node));
142 15             new_doc->doc_id = doc_id;
143 16             new_doc->next = current->docs_list;
144 17             current->docs_list = new_doc;
145 18             return;
146 19         }
147 20         current = current->next;
148 21     }
149 22
150 23     // Si la palabra no se encuentra en el índice, crea un nuevo nodo en el índice.
151 24     InvertedIndex *new_node = create_new_node(word);
152 25     new_node->next = hash_table[index]; // Posiciona el nuevo nodo en el índice calculado.
153 26     hash_table[index] = new_node;
154 27
155 28     // Crea y asigna el nodo de documento.
156 29     Node *new_doc = (Node *)malloc(sizeof(Node));
157 30     if (new_doc == NULL)
158 31     {
159 32         fprintf(stderr, "Error al asignar memoria para nuevo documento\n");
160 33         exit(EXIT_FAILURE);
161 34     }
162 35     new_doc->doc_id = doc_id;
163 36     new_doc->next = NULL;
164 37     new_node->docs_list = new_doc;
165 38 }
166 39

```

**Código 7.** Función para añadir un nuevo documento al índice invertido.

Se visitó [GeeksforGeeks](#) para tener información sobre el índice invertido [2].

### 4.3. Conexión de las aristas con los vértices

En el grafo dirigido, los documentos se representan como vértices y los enlaces entre ellos como aristas. Para gestionar estos enlaces, se utilizan listas de adyacencia, donde cada documento tiene una lista que contiene los documentos a los que enlaza (enlaces salientes) y una lista de los documentos que lo enlazan (enlaces entrantes).

La función `add_edge()` es la encargada de agregar un enlace entre dos documentos en el grafo. Esta función realiza las siguientes acciones:

- Primero, se crea un nodo en la lista de adyacencia de salientes para el documento de origen. Este nodo almacena el ID del documento de destino al que el documento de origen enlaza.

- Luego, se crea un nodo en la lista de adyacencia de entrantes para el documento de destino. Este nodo almacena el ID del documento de origen desde el cual el documento de destino recibe un enlace.

Ambos nodos se insertan al principio de sus respectivas listas, de manera que cada documento tiene una lista que enlaza a los documentos a los que hace referencia y una lista que lo vincula con los documentos que lo referencian. Esto permite un acceso rápido a los documentos relacionados con cualquier documento dado en el grafo.

En la inicialización del grafo, las listas de adyacencia de salientes y entrantes se inicializan a NULL, lo que indica que no hay enlaces. Posteriormente, cuando se procesa un archivo que contiene enlaces, se agregan los nodos correspondientes a las listas de adyacencia de los documentos implicados.

Es importante destacar que, cuando se agrega un nuevo enlace, se verifica si se puede asignar memoria para el nuevo nodo. Si la asignación de memoria falla, se imprime un mensaje de error y el programa termina.

Las aristas en el grafo se conectan a los vértices mediante las listas de adyacencia de salientes y entrantes. Cada vez que se encuentra un enlace en un archivo, se actualizan las listas de adyacencia correspondientes para reflejar la relación entre los documentos, permitiendo la navegación eficiente entre ellos.

```

1 // Agrega un enlace dirigido entre dos documentos en el grafo.
2 void add_edge(Graph *graph, int source, int destination)
3 {
4     // Crea nodo en la lista de enlaces salientes desde el documento de origen al destino.
5     Node *newOutputNode = (Node *)malloc(sizeof(Node));
6     if (newOutputNode == NULL) // Verifica la asignación de memoria.
7     {
8         printf("Error al asignar memoria para enlace saliente\n");
9         exit(EXIT_FAILURE);
10    }
11    newOutputNode->doc_id = destination; // Asigna el ID de destino al nodo.
12    newOutputNode->next = graph->output_adjacent_list[source]; // Inserta al inicio de la lista
13    graph->output_adjacent_list[source] = newOutputNode;
14
15    // Crea nodo en la lista de enlaces entrantes al documento destino desde el origen.
16    Node *newInputNode = (Node *)malloc(sizeof(Node));
17    if (newInputNode == NULL) // Verifica la asignación de memoria.
18    {
19        printf("Error al asignar memoria para enlace entrante\n");
20        exit(EXIT_FAILURE);
21    }
22    newInputNode->doc_id = source; // Asigna el ID del origen al nodo.
23    newInputNode->next = graph->input_adjacent_list[destination]; // Inserta al inicio de la lista.
24    graph->input_adjacent_list[destination] = newInputNode;
25 }

```

**Código 8.** Función para añadir un enlace entre dos documentos.

#### 4.4. Creación de archivos

Se generaron los archivos de texto que simulan documentos web, con contenido aleatorio y enlaces a otros documentos dentro de la misma carpeta. Cada archivo creado representa un documento que podría pertenecer a un grafo web. La función principal para esta tarea es `generate_text_files()`, la cual se encarga de crear múltiples archivos, mientras que `generate_random_text()` se utiliza para rellenar cada archivo con contenido aleatorio y enlaces.

La función `generate_text_files()` recibe como parámetro el número de documentos a generar y realiza los siguientes pasos:

- Primero, verifica que el número de documentos sea positivo. Si el valor es menor o igual a cero, imprime un mensaje de error en la salida estándar de error y termina el programa, ya que al menos un archivo debe generarse.
- A continuación, por medio de un ciclo, verifica que cada documento tenga al menos un enlace.
- Luego, para cada documento a crear, genera un nombre único en el formato `doc<i>.txt` (por ejemplo, `doc1.txt`, `doc2.txt`, etc.), donde `i` es el número del documento.
- Intenta abrir cada archivo en modo de escritura. Si no puede abrir el archivo, imprime un mensaje de error en la salida estándar de error y termina el programa.
- Finalmente, llama a la función `generate_random_text()` para llenar el archivo con texto y enlaces, y luego cierra el archivo.

La función `generate_random_text()` tiene la responsabilidad de escribir el contenido en cada archivo, recibe como parámetro el número de caracteres a generar y realiza las siguientes acciones:

- Primero, verifica que el número de caracteres sea positivo. Si el valor es menor o igual a cero, imprime un mensaje de error en la salida estándar de error y termina el programa, ya que debe generarse al menos un carácter.
- Después, Escribe letras aleatorias en el archivo, generando un total de `num_characters` caracteres, que es el ingresado anteriormente. Para cada carácter, se selecciona una letra mayúscula aleatoria entre 'A' y 'Z', y se escribe en el archivo. Entre cada letra se agrega un espacio, lo que simula contenido de texto.

- Luego, genera un número aleatorio de enlaces en relación con la cantidad de archivos generados, hacia otros documentos del conjunto. Para cada enlace, selecciona un número de documento de destino de forma aleatoria. Si el documento seleccionado no es el mismo documento en el que se está escribiendo, se añade una línea en el formato link: doc<i> para representar un enlace a otro documento.

```

196 1
197 2
198 3 // Genera texto aleatorio y enlaces en un archivo web específico.
199 4 void generate_random_text(FILE *doc, const char *doc_name, int num_docs, int num_characters, int
200 5     current_doc, int *links)
201 6 {
202 7     // Genera texto aleatorio dentro de cada archivo web.
203 8     for (int i = 0; i < num_characters; i++)
204 9     {
205 10         char letter = 'A' + rand() % 26; // Genera una letra aleatoria entre 'A' y 'Z'.
206 11         fprintf(doc, "%c", letter);      // Escribe la letra en el archivo.
207 12         if (i < num_characters - 1)
208 13             fprintf(doc, " "); // Agrega un espacio entre las letras.
209 14     }
210 15
211 16     fprintf(doc, "\nlink: doc%d", links[current_doc - 1]); // Asegura al menos un enlace nico .
212 17
213 18     // Incluye aleatoriamente enlaces a otros documentos.
214 19     int extra_links = rand() % num_docs; // Genera enlaces respecto al numero de documentos.
215 20     for (int i = 0; i < extra_links; i++)
216 21     {
217 22         int link_doc = rand() % num_docs + 1; // Elige un documento
218 23         aleatorio de 1 a num_docs.
219 24         if (link_doc != current_doc && link_doc != links[current_doc - 1]) // Evita que un documento
220 25         enlace a s mismo.
221 26             fprintf(doc, "\nlink: doc%d", link_doc); // Agrega un enlace al
222 27         archivo.
223 28     }
224 29
225 30     fprintf(stdout, "ARCHIVO '%s' generado con XITO con %d letras y enlaces.\n", doc_name,
226 31         num_characters);
227 32 }
228 33

```

**Código 9.** Función de generación de texto random.



```

1 void generate_text_files(int num_docs, int num_characters)
2 {
3     if (num_docs <= 0 || num_docs >= 100) // Verifica que se proporcione al menos un documento.
4     {
5         fprintf(stderr, "El n mero de archivos web a generar debe ser MAYOR a 0 y MENOR a 100.\n\n");
6         exit(EXIT_FAILURE);
7     }
8
9     // Array para asegurar que cada documento tenga al menos un enlace.
10    int *links = malloc(num_docs * sizeof(int));
11
12    for (int i = 0; i < num_docs; i++)
13        links[i] = (i + 1) % num_docs + 1; // Crea un ciclo simple para asegurar al menos un enlace por
14        archivo.
15
16    // Itera para crear el n mero especificado de archivos de texto.
17    for (int i = 1; i <= num_docs; i++)
18    {
19        char doc_name[MAX_NAME_DOC];
20        snprintf(doc_name, sizeof(doc_name), "doc%d.txt", i); // Genera un nombre para el archivo (e.g.,
21        "doc1.txt").
22
23        // Crea el archivo en modo escritura. // Termina el programa si no se puede abrir el archivo.
24        FILE *doc = fopen(doc_name, "w");
25        if (doc == NULL)
26        {
27            fprintf(stderr, "Error al abrir el archivo web.\n");
28            free(links);
29            exit(EXIT_FAILURE);
30        }
31
32        // Genera contenido aleatorio y enlaces dentro del archivo.
33        generate_random_text(doc, doc_name, num_docs, num_characters, i, links);
34
35        fclose(doc);
36    }
37 }

```

Código 10. Función de generación de archivos.

## 5. Hilo de ejecución y salida del programa

### Recorrido del programa

Para poder ver todo el funcionamiento del programa, puede visitar el repositorio en *GitHub*, acá se detallarán las cosas mas relevantes para dar un vistazo general del proyecto.

### 5.1. Entrada del programa

Para esta parte hay que considerar que el programa se divide en dos carpetas *src*, una que contiene los archivos para la creación de documentos(*src2*) y otra que contiene los archivos para el simulador de búsqueda(*src*).

#### 5.1.1. *src2* Creación de Archivos

El programa comienza con la inclusión de los archivos de cabecera *graph.h* y *doc.h*, que definen las estructuras y funciones necesarias para el proceso de creación de archivos. A continuación, el *main()* de *src2* gestiona los argumentos proporcionados por la línea de comandos, utilizando la función *getopt()* para manejar las opciones disponibles. El programa acepta tres opciones principales:

- **-h**: Muestra un mensaje de ayuda que explica cómo utilizar el programa, indicando que se debe usar la opción **-d** y **-c** seguidas de un número para especificar la cantidad de archivos a generar y la cantidad de caracteres por archivo.
- **-d**: Permite al usuario especificar la cantidad de documentos a procesar. El valor se recibe como un argumento entero y se almacena en la variable *num\_docs*. Si no se especifica un valor o si el valor es menor o igual a cero, el programa mostrará un mensaje de error y finalizará.
- **-c**: Permite al usuario definir la cantidad de caracteres que se generarán en cada archivo. El valor se recibe como un argumento entero y se almacena en la variable *num\_characters*. Si no se proporciona un valor o si el valor es menor o igual a cero, el programa mostrará un mensaje de error y finalizará.

Si el usuario no proporciona los parámetros adecuados, se mostrarán mensajes de error, ya sea por una opción no reconocida o un valor inválido para el número de documentos.

El flujo del programa es el siguiente:

- Primero, se inicializan las variables y estructuras necesarias.
- Luego, se llama a la función *generate\_text\_files()* para crear los archivos de texto según la cantidad de documentos especificada.
- Dentro de cada archivo generado, se invoca la función *generate\_random\_text()* para generar el contenido de caracteres.
- Finalmente, se imprime un mensaje de confirmación indicando la creación exitosa de los archivos.

A continuación, se muestra el fragmento relevante del código que gestiona la entrada de los parámetros:

```

268 1  int opt;
269 2  int num_docs = 0;
270 3  int num_characters = 0;
271 4
272 5  while ((opt = getopt(argc, argv, "hd:c:")) != -1)
273 6  {
274 7      switch (opt)
275 8      {
276 9          case 'h':
277 10             fprintf(stdout, "\nUso del programa:\n");
278 11             fprintf(stdout, "  -d <numero_de_archivos> : N mero de archivos a crear.\n");
279 12             fprintf(stdout, "  -c <numero_de_caracteres> : N mero de caracteres por archivo.\n");
280 13             fprintf(stdout, "  -h : Muestra esta ayuda.\n\n");
281 14             exit(EXIT_SUCCESS);
282 15             break;
283 16          case 'd':
284 17             num_docs = atoi(optarg);
285 18             break;
286 19          case 'c':
287 20             num_characters = atoi(optarg);
288 21             break;
289 22          case '?':
290 23             fprintf(stderr, "Opci n no reconocida: -%c\n", optopt);
291 24             exit(EXIT_FAILURE);
292 25          default:
293 26             fprintf(stderr, "Uso: %s [-h] [-d numero_de_archivos] [-c numero_de_caracteres]\n", argv[0])
294 27             ;
295 28             exit(EXIT_FAILURE);
296 29         }
297
298     }
299

```

**Código 11.** Función que maneja la entrada en *src2*.

### 5.1.2. src Simulador de búsqueda

El programa comienza con la inclusión de los archivos de cabeceras, `doc.h`, `graph.h`, `inverted_index.h` y `pagerank.h` que definen las estructuras y funciones necesarias para el funcionamiento del algoritmo. A continuación, el `main()` de `src` procesa los argumentos de línea de comandos, utilizando la función `getopt()` para gestionar las opciones disponibles. El programa acepta dos opciones principales:

- **-h**: Muestra un mensaje de ayuda que explica cómo usar el programa, indicando que se debe usar la opción **-s** seguida de una palabra para que se pueda buscar.
- **-s**: Permite que el usuario ingrese una palabra para buscar.

Si el usuario no proporciona los parámetros adecuados, se muestran mensajes de error, como en el caso de una opción no reconocida o un valor inválido para el número de documentos.

El flujo del programa es el siguiente:

- Primero, se inicializan variables y estructuras necesarias, como el arreglo `pagerank` para almacenar los valores de PageRank y la estructura `Graph` para representar el grafo.
- Posteriormente, se construye el grafo mediante las funciones `initialize_graph()`, `build_graph()`, `build_index()`, y se imprime la palabra buscada con `print_search_word()`.
- Finalmente, se calcula y muestra el PageRank utilizando `calculate_pagerank()` y `display_pagerank()`, para luego liberar los recursos con `release_graph()`.

A continuación, se muestra el fragmento relevante del código que gestiona la entrada de los parámetros:

```

1  int opt;
2  char *word_to_search = NULL;
3
4  while ((opt = getopt(argc, argv, "hs:")) != -1)
5  {
6      switch (opt)
7      {
8          case 'h':
9              fprintf(stdout, "\nPara ingresar la palabra a buscar, por favor coloque el par metro <-s> <
10             numero_de_archivos>\n\n");
11             break;
12          case 's':
13             word_to_search = optarg;
14             break;
15          case '?':
16             fprintf(stderr, "Opci n no reconocida: -%c\n", optopt);
17             exit(EXIT_FAILURE);
18          default:
19             fprintf(stderr, "Uso: %s [-h] [-s palabra_a_buscar]\n", argv[0]);
20             exit(EXIT_FAILURE);
21     }
22 }

```

**Código 12.** Función que maneja la entrada en `src`.

## 5.2. Salida del programa

En la imagen se muestra un ejemplo de la salida generada por el programa. A continuación, se describe cada parte de esta salida:

```

1  miguel@miguel:~/Escritorio/search-simulator$ make docs
2
3  Número de archivos: 3
4
5  ARCHIVO 'doc1.txt' generado con ÉXITO con 50 letras y enlaces.
6  ARCHIVO 'doc2.txt' generado con ÉXITO con 50 letras y enlaces.
7  ARCHIVO 'doc3.txt' generado con ÉXITO con 50 letras y enlaces.
8
9  miguel@miguel:~/Escritorio/search-simulator$ make run
10
11 Valores de PageRank:
12 Documento 0 (doc3.txt): PageRank = 0.050000
13 Documento 1 (doc1.txt): PageRank = 0.151184
14 Documento 2 (doc2.txt): PageRank = 0.178561
15 Grafo de enlaces:
16 Documento 1 enlaza a: 2
17 Documento 2 enlaza a: 1 1 3
18 Documento 3 enlaza a: 2 2 1
19 La palabra 'f' se encuentra en los siguientes documentos:
20 Doc1: 1 veces
21 Doc3: 1 veces
22 Archivo EPS generado: graph.eps

```

Figura 1. Salida del programa.

- **Generación de archivos:** El programa inicia generando el número de archivos especificado (en este caso, 3). Para cada archivo, se indica su nombre, su tamaño (en número de letras) y las conexiones generadas entre ellos.
- **Cálculo del PageRank:** Se imprimen los valores de PageRank de cada documento. Estos valores reflejan la relevancia de cada documento en función de las conexiones entre ellos.
- **Grafo de enlaces:** Se detalla cómo están conectados los documentos entre sí. Por ejemplo, el documento 1 enlaza a los documentos 2 y 3, mientras que el documento 3 enlaza al 1.
- **Índice invertido:** Se muestra en qué documentos se encuentra una palabra específica (en este caso, la palabra f). Por ejemplo, esta palabra aparece una vez en los documentos doc1.txt y doc3.txt.
- **Archivo EPS:** Finalmente, se genera un archivo en formato .eps que contiene una representación gráfica del grafo de enlaces.

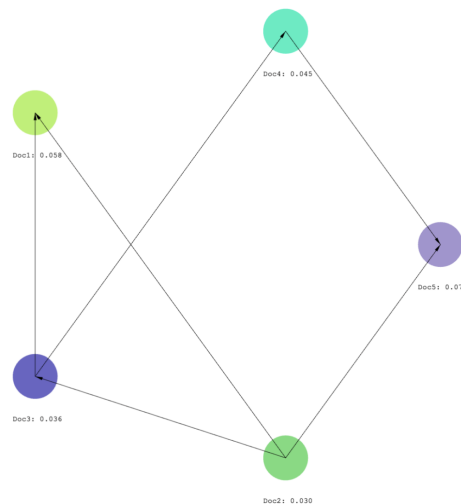


Figura 2. Representación gráfica de los grafos.

## 6. Dificultades enfrentadas

### 6.1. Hasheo y tokenización

Una de las dificultades en el desarrollo fue implementar un sistema eficiente de búsqueda y acceso a las palabras clave en los documentos. Para reducir el tiempo de búsqueda y evitar una complejidad de  $O(n)$  en cada comparación de palabras, se implementó una tabla hash que permitiera acceder a cada palabra en tiempo constante,  $O(1)$ , en promedio.

La mayor complicación fue el proceso de construir el hash de cada palabra, el cual implicaba una serie de pasos. Primero, se requería tokenizar el texto, es decir, dividirlo en palabras individuales. Esto era un reto en sí mismo, ya que había que ignorar las "stopwords" (palabras o letras comunes, que no aportan valor al análisis) para evitar almacenar términos irrelevantes en el hash. Además, fue necesario implementar un filtro para eliminar los símbolos y signos de puntuación que no eran útiles, dejando solo las palabras significativas para el cálculo del hash.

Estos pasos adicionales agregaron dificultad al proceso, pero fueron importantes para lograr una estructura de datos eficiente y optimizada para consultas rápidas.

### 6.2. Generación de la Imagen EPS

Al intentar crear la imagen en formato EPS del grafo, se observó que, aunque los nodos se generan correctamente, los enlaces entre ellos no se visualizan.

### 6.3. Separación de archivos

Durante la organización del trabajo, se generaron dos directorios principales para el código fuente: `src` y `src2`. En el directorio `src2`, además de incluir nuevos archivos fuente para funciones adicionales, se creó un segundo archivo `main.c`, lo que causó problemas durante el proceso de compilación.

El problema radica en que, al compilar el proyecto, el compilador encontró dos puntos de entrada al programa (los archivos `main.c` en `src` y `src2`). Dado que en C solo puede existir un único punto de entrada definido por la función `main()`, se generaron errores de compilación.

La estructura del proyecto que causó este conflicto es la siguiente:

```

src
  graph.c
  graphic.c
  inverted_index.c
  main.c
  pagerank.c
src2
  doc.c
  main.c

```

El archivo `main.c` en `src2` introdujo el error al estar presente otro `main.c` en `src`. Como resultado, el compilador generó errores indicando múltiples definiciones de la función `main()`.

## 7. Conclusión

Se realizó con éxito un sistema de recuperación de información, implementando cada uno de los requerimientos entregados en el PDF. Al realizar esta simulación, se pudo observar la importancia de tener algoritmos eficientes y métodos de búsqueda efectivos, acercándose a  $O(1)$  en el caso ideal.

Este ejercicio nos muestra la necesidad de optimizar no solo la velocidad de los algoritmos, sino también el uso de la memoria y los recursos. Nos permite comprender mejor cómo diferentes estructuras de datos pueden influir en el rendimiento de un sistema, así como los factores que se deben considerar al seleccionar o diseñar un algoritmo.

Además, esta implementación demuestra la relevancia de construir estructuras de datos adecuadas para problemas específicos, lo que nos prepara para abordar proyectos más complejos en el futuro.

Tener en cuenta aspectos como la eficiencia y la escalabilidad se convierte en una habilidad importante para desarrollar soluciones robustas y de alto rendimiento.

## ■ Referencias

- [1] W. contributors. «PageRank». Último acceso: 18 de noviembre de 2024. (2024), dirección: <https://en.wikipedia.org/wiki/PageRank>.
- [2] GeeksforGeeks. «Inverted Index». Último acceso: 18 de noviembre de 2024. (2024), dirección: <https://www.geeksforgeeks.org/inverted-index/>.
- [3] GeeksforGeeks. «PageRank Algorithm Implementation». Último acceso: 18 de noviembre de 2024. (2024), dirección: <https://www.geeksforgeeks.org/page-rank-algorithm-implementation/>.
- [4] Memgraph. «PageRank Algorithm». Último acceso: 18 de noviembre de 2024. (2024), dirección: <https://memgraph.com/docs/advanced-algorithms/available-algorithms/pagerank>.