Universidad de Chile
Facultad de Ciencias Físicas y Matemáticas
Depto. de Ciencias de la Computación
CC4102 - Diseño y Análisis de Algoritmos

Tarea 2

Integrantes : Americo Ferrada
Belisario Panay
Profesor : Pablo Barcelo
Ayudantes : Claudio Torres
Jaime Salas
Auxiliar : Ariel Cáceres

Índice

1. Introducción

En el presente informe se muestra el diseño, implementación y experimentación de dos diferentes enfoques para la búsqueda de texto, estos son el enfoque basado en arreglo de sufijos y el otro es el basado en el algoritmo con autómata.

En particular el arreglo de sufijos es un arreglo de enteros, donde cada entero apunta a un caracter del texto, el cual representa un sufijo de este, estos sufijos están ordenados lexicograficamente en el arreglo. Para el caso del automata es una máquina de estados para aceptar al patrón buscado, este automata es ejecutado sobre el texto.

La idea de la tarea es que ambos algoritmos tomen $O(n + m)$ en buscar las posiciones en que se puede encontrar un patrón de largo (m) en un texto de largo n, siendo la diferencia entre los dos enfoques que para el arreglo de sufijos toma $O(n)$ en construir el arreglo y $O(m)$ para encontrar los apariciones del patrón, en cambio para el algoritmo del automata toma tiempo $O(m)$ construir el automata y $O(n)$ correrlo en el texto para encontrar las apariciones del patrón.

1.1. Problema a resolver

Un algoritmo estándar de creación de suffix array toma tiempo $O(n^2 \log n)$, mientras que el algoritmo que se implementará toma tiempo $O(n)$ para su construcción. Para la creación del automata el algotimo más simple es $O(m^3)$, y el que se implementará usará tiempo $O(m)$. El problema consiste en realizar los pasos necesarios para la buena implementación del algoritmo, ya que pequeños errores en código pueden producir un algoritmo de mayor orden de magnitud y con ello se falla en el objetivo.
Los pasos para realizarlos estan detallados en el enunciado de la tarea (archivos adjuntos) y el *paper* de Juha Karkkaneinen y Peter Sanders, creadores del algoritmo para la construcción del suffix array en tiempo lieal.

1.2. Hipótesis

Se espera que la implementacion en Java no afecte de manera notoria el tiempo de ejecución de los algoritmos. Para la construccion del suffix array se espera que ocupe un tiempo lineal para su construcción, y que la constante que lo acompaña la cual se espera que sea el tamaño del alfabeto (debido a la sub-rutina de radix sort) no afecte de manera considerable los tiempos de ejecucuion del algoritmo, para la busqueda del patrón en el suffix array se espera obtener el $O(m \log n)$ mostrado en el enunciado de la tarea.

En el caso del automata se espera que la construcción del automata tome tiempo del orden del tamaño del patrón, no afectando de manera notoria la constante que lo acompaña que se espera sea el tamaño del alfabeto. Para ejecutarlo sobre texto se necesitará del orden del tamaño del texto.

2. Diseño Teórico

Para los experimentos se pidió hacer mediciones de los tiempos de distintos tamaños de textos, para tomar en cuenta estos tamaños se tomaron una serie de consideraciones. Se tomó en cuenta que luego de pre-procesar un texto, disminuir su tamaño aprox 0.788 del tamaño original. Se tomó como estimación del tamaño que las palabras tienen largo 6. Con esto quedaron los siguientes tamaños:

- 250 Kbs 2^15
- 500 Kbs 2^16
- 1 Mbs 2^17
- 2 Mbs 2^18
- 4 Mbs 2^19
- 8 Mbs 2^20
- 16 Mbs 2^21
- 32 Mbs 2^22
- 64 Mbs 2^23
- 128 Mbs 2^24
- 256 Mbs 2^25

2.1. Main

En Main se crea un archivo de *logging* para registrar el tiempo de creación del Suffix Tree y los resultados de búsqueda, a partir de un texto leído desde el disco. En particular:

- El texto leído se preprocesa (se eliminan puntuaciones, espacios, saltos de línea y todo lo que no corresponda al *regex* [a-ZA-Z]).
- Se crea el Suffix Tree usando el algoritmo de Ukkonen.
- Se registra el tiempo de creación.
- Se generan palabras aleatorias del texto, para buscarlas usando el Suffix Tree.
- Se registran los resultados de búsqueda.

2.2. Ukkonen

Clase principal del algoritmo. Aquí se realiza la totalidad de la ejecución del algoritmo de Ukkonen.
Posee 5 métodos auxiliares:

- run(): Crea el Suffix Tree correspondiente, usando los siguientes métodos auxiliares.
- getPath(char s, Node n): Retorna el camino desde el nodo n, con el caracter s.
- search(String suffix, Node root): Busca el String *suffix* en el nodo *root*
- getSuffixes(Node root, String suffix, int count): Método auxiliar para search. Busca recursivamente en los nodos por el sufijo *suffix* usando el número *count* (usado para los caminos).
- getLeafPath(Node n, int count): Retorna el camino a partir del nodo n, recorriendo *count* nodos internos.

2.3. Last

Esta clase es usada para mostrar el final de un camino.

2.4. Node

Esta clase se usa para almacenar la información correspondiente a cada sufijo. Puede ser o no ser una hoja. Es clase padre de *InternalNode*.

2.5. InternalNode

Esta clase extiende de *Node* para usarse como nodo interno (i.e., explícitamente **no es** una hoja).

2.6. TextPreprocessor

Esta clase recibe un texto como String y devuelve solamente los caracteres que coinciden con el *regex* [a-zA-Z]. El resto de los caracteres se eliminan (entre los cuales están los saltos de línea, las puntuaciones y los espacios)

2.7. Logger

Esta clase recibe un nombre de archivo y escribe los datos que se le entregan a dicho archivo. Sirve para registrar la información necesaria para generar los gráficos del informe.

3. Presentación de los Resultados

3.1. Tiempo de Creación del Suffix Tree

Los resultados para los tiempos de construcción del Suffix Tree usando nuestra implementación son los siguientes:

Largo del texto	Tiempo de Construcción (mseg)
2^{15}	5
2^{16}	3
2^{17}	3
2^{18}	6
2^{19}	11
2^{20}	13
2^{21}	21
2^{22}	44
2^{23}	93
2^{24}	167
2^{25}	329

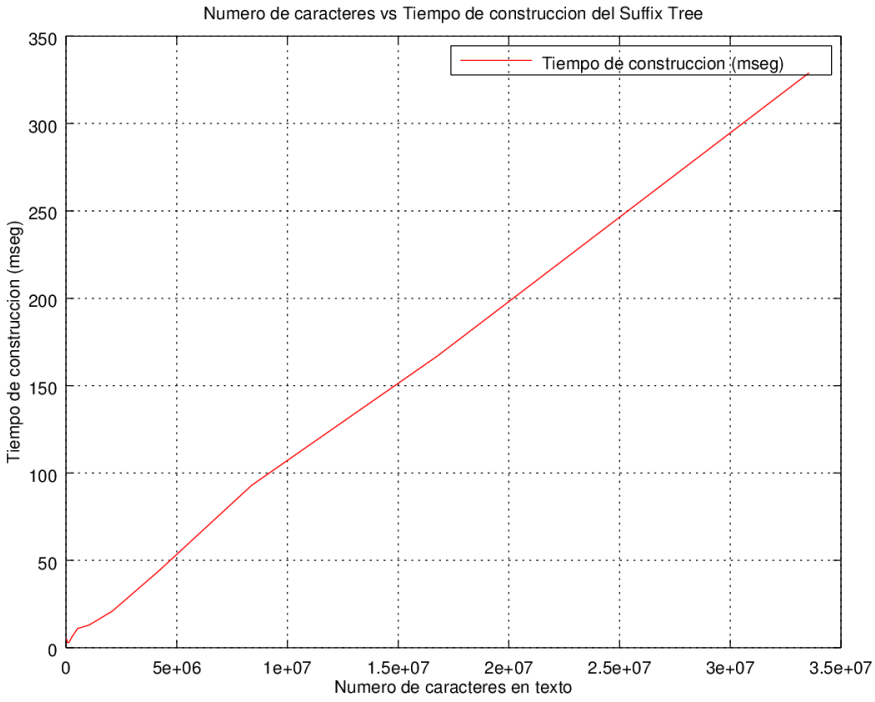

Figura 1: Tiempo de creación del Suffix Tree para N = $2^{15.,25}$

3.2. Desempeño de operación *Buscar*

Los resultados para los tiempos de búsqueda en el Suffix Tree son los siguientes:

Número de palabras	Largo promedio del patrón	Tiempo de búsqueda (nano-segundos)
2^{15}	5.36	2671.75
2^{16}	5.31	1109.40
2^{17}	5.52	259.94
2^{18}	5.39	296.97
2^{19}	5.35	254.40
2^{20}	4.76	268.31
2^{21}	4.50	246.81
2^{22}	4.43	245.39
2^{23}	4.37	254.88
2^{24}	4.42	253.84
2^{25}	4.47	261.40

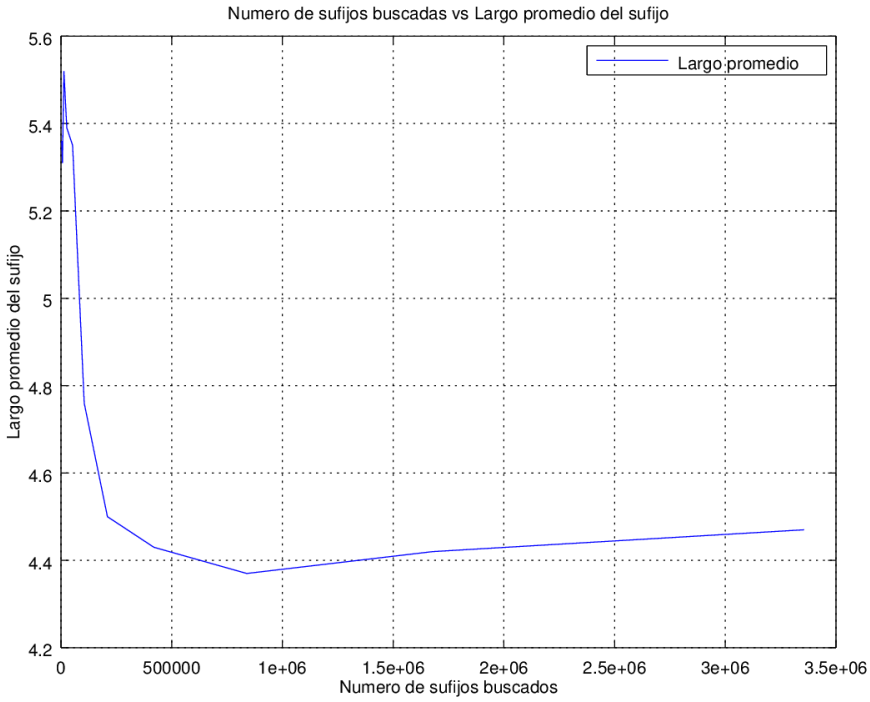

Figura 2: Largo promedio del patrón

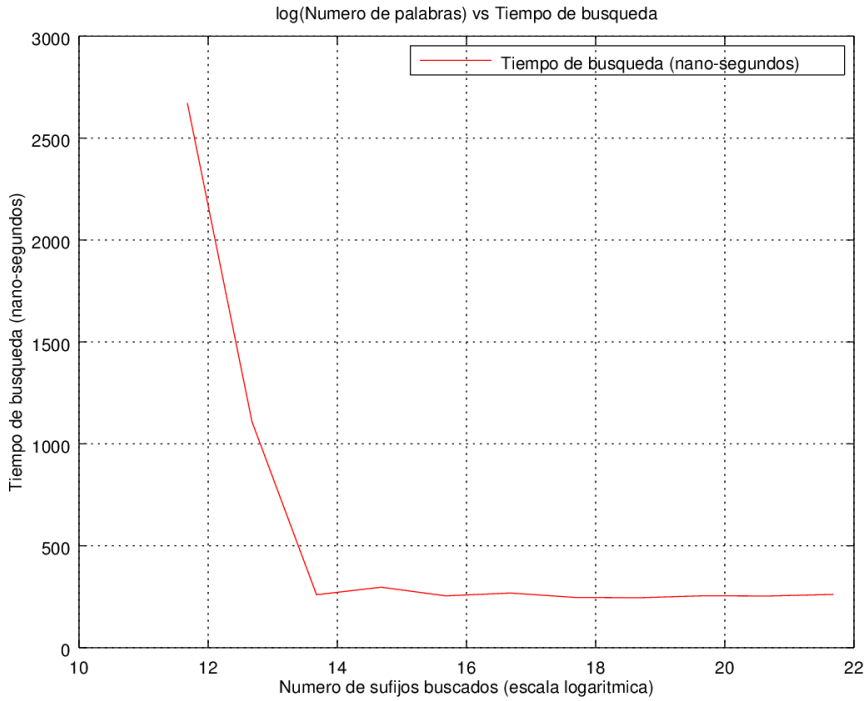

Figura 3: Tiempo de búsqueda

4. Análisis y Conclusiones

4.1. Construcción del Suffix Tree

Nuestra implementación funciona bien cuando se usan palabras cortas, dado que es posible construir el Suffix Tree en tiempo $O(n)$ (puesto que al duplicar el largo de la palabra, también se duplicaba el tiempo en nano-segundos para construirlo).
Nuestra hipótesis de que el *garbage collector* de Java interferiría con nuestros resultados es correcta, dado que cada vez que se corría el algoritmo para una cantidad fija de caracteres variaba notablemente. Aún así, su interferencia es muy baja como para considerarla válida, por lo que solo lo explicamos por motivos de completitud.
Tomando este caso, encontramos que el algoritmo de Ukkonen tarda (por ejemplo) 5 milisegundos para $N = 2^{15}$ caracteres, y se va duplicando (aproximadamente) cada vez que duplicamos la cantidad de caracteres, por lo que la implementación parece haber sido conseguida en tiempo $O(n)$.
A último momento, nos dimos cuenta que nuestra implementación tenía problemas al momento de ejecutarse con textos largos, dado que no creaba correctamente los nodos internos. De esta forma el arbol se creaba demasiado rápido como para ser coherente con la cantidad de enlaces que deberían existir en el Suffix Tree.

4.2. Búsqueda en el Suffix Tree

Encontramos que la búsqueda en el Suffix Tree (respecto a nuestra implementación) cae notablemente, lo que probablemente se debe a que mientras más sufijos se tienen, más sufijos se descartan al momento de buscar la respuesta final. Notar que en nuestra implementación, un **break** en el código *bypassea* el *loop*, cuando en el *loop* se debería salir solo cuando no quedan más sufijos por procesar.
En resumen, mientras más sufijos hay, más quedan sin procesar al momento de buscar, por lo que la cantidad de pasadas es menor.

4.3. Conclusiones

Creemos que el problema en la creación de nodos internos se debe a la creación errónea de los *suffix links* o al seguimiento incorrecto de éstos. Las funciones de conteo están correctas para Suffix Trees bien construídos (i.e. de tamaño pequeño en nuestra implementación), pero pueden mejorarse debido a que se usa recursión para su implementación. Si hubieramos construído correctamente el Suffix Tree con los textos extensos de prueba, probablemente lanzaría un *Stack Overflow*, por lo que en ese caso se cambiaría a una implementación iterativa.