

Tarea 3: Skip Lists y ABB Aleatorizados

CC4102 - Diseño y Análisis de Algoritmos

Profesor:	Gonzalo Navarro
Auxiliar:	Teresa Bracamonte
Alumnos:	Cristián Carreño
	Sergio Maass
Fecha:	18 de Diciembre de 2013

Índice

1. Introducción	2
2. Implementación	2
2.1. Skip List	2
2.1.1. Clase SkipNode	3
2.1.2. Clase SkipList	3
2.2. ABB	4
2.3. ABB Aleatorizado	4
3. Experimentos	6
3.1. Tiempos de inserción y búsqueda	7
3.1.1. Inserción en ABB aleatorizado	7
3.1.2. Inserción en Skip Lists	8
3.1.3. Búsqueda en ABB aleatorizado	9
3.1.4. Búsqueda en Skip Lists	10
3.2. Comparación entre ABB y ABB aleatorizado	11
4. Conclusiones	12

1. Introducción

Los Skip Lists (SL) son estructuras de datos que permiten almacenar objetos en listas enlazadas ordenadas. A diferencia de las listas enlazadas convencionales, los skip lists aleatorizan el algoritmo de inserción de forma que cada vez que se inserta un nuevo elemento, se pueden generar d duplicados aleatoriamente de forma que estos compongan un atajo para la lista enlazada original y reduzcan el tiempo de búsqueda. La ventaja de los skip list es que aseguran un tiempo promedio de búsqueda de $O(\log n)$, al igual que los árboles de búsqueda binarios.

Los Árboles de Búsqueda Binarios (ABB) aseguran que, en promedio, los tiempo de inserción y búsqueda sean de $O(\log n)$. Sin embargo, en el peor caso podemos terminar con un árbol de altura $O(n)$. Para reducir ese tipo de situaciones, Martinez y Roura proponen aleatorizar las operaciones de inserción y eliminación en el ABB. Específicamente, en la inserción se considera que cada elemento tiene la probabilidad $\frac{1}{n}$ de ser la raíz del árbol; siendo n el número de elementos actualmente en el árbol.

En el presente trabajo se pretende determinar cuál de estas estructuras es mejor en la práctica, para lo cual se implementarán y luego serán sometidas a prueba.

2. Implementación

A continuación se detalla la implementación de las estructuras que serán comparadas: Skip List (SL), Árbol Binario de Búsqueda (ABB), y ABB Aleatorizado. Todas se implementaron en Python.

2.1. Skip List

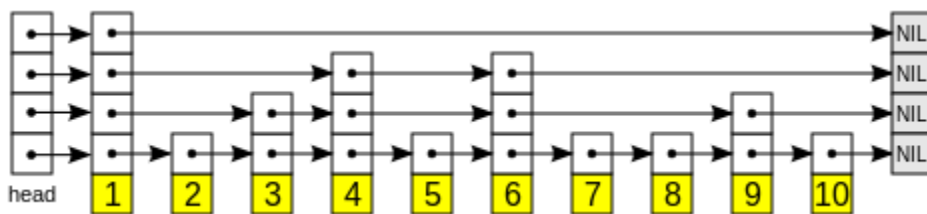


Figura 1: Estructura de Skip List.

La estructura de una Skip List se aprecia en la figura 1. La altura de cada nodo (cantidad de punteros que tiene en listas superiores) se determina del siguiente modo: se elige con probabilidad $\frac{1}{2}$ si el nodo tiene un puntero en el nivel i (comenzando con $i = 2$), y en caso afirmativo se procede con la elección para el nivel $i + 1$, en caso negativo se termina el proceso. Todo nodo tiene al menos altura 1 (está presente en la primera lista).

Para implementar esta estructura se definieron las clases SkipNode y SkipList.

2.1.1. Clase SkipNode

Representa a cada nodo y sus duplicados en listas superiores. Consiste simplemente en una llave *key* y una lista *next* de tamaño n , donde para cada $i \in [0, n)$, $next_i$ es un puntero al siguiente elemento de la i -ésima lista de la Skip List.

2.1.2. Clase SkipList

Representa a la lista en sí, la cual está conformada por un SkipNode *head* y las operaciones de inserción y búsqueda.

Ambas operaciones hacen uso del método *closest_nodes(k)*, el cual entrega una lista de tamaño igual a la altura de la Skip List, y que contiene punteros a los nodos cuyas llaves tienen los valores más cercanos a k por la izquierda (con valor menor a k) en cada nivel. El método funciona del siguiente modo:

closest_nodes(k):

- Se crea una lista u de tamaño igual a la altura de la Skip List, inicializada a *None*.
- Se asigna $node := head$.
- Se itera i desde $|u|$ a 1:
 - Mientras $node.next[i] \neq None$ y $node.next[i].key < k$
 - $node := node.next[i]$
 - $u[i] := node$
- Retorna u

Luego, podemos definir la inserción del siguiente modo:

insert(k):

- Se crea un SkipNode *node* con la llave k que se desea insertar, y una altura determinada con el método explicado en 2.1.
- Si la altura de *node* es mayor a la de *head*, se añaden a *head* los niveles faltantes (con *None*).
- Se obtiene la lista de nodos más cercanos por la izquierda en cada nivel, $u := closest_nodes(k)$
- Para cada nivel i de la altura de *node*:
 - $node.next[i] := u[i].next[i]$
 - $u[i].next[i] := node$

Finalmente, la búsqueda queda:

search(k):

- Se obtiene la lista de nodos más cercanos al valor de k por la izquierda en cada nivel, $u := \text{closest_nodes}(k)$.
- Si la longitud de u es mayor que cero:
 - Se asigna $c := u[0].\text{next}[0]$ (si el nodo cuya llave es k está en el árbol, ahora será apuntado por c).
 - Si $c \neq \text{None}$ y $c.\text{key} = k$, c es el nodo buscado y por ende se retorna.
- Si se llegó a este punto la búsqueda fue infructuosa y se retorna *None*.

2.2. ABB

El ABB se implementó con una clase *Node* y una clase *ABB*. Cada objeto *Node* contiene una llave *key* y los punteros a sus hijos *left* y *right*. Una instancia de *ABB* contiene a un objeto *Node root* y las operaciones *insert* y *search*. Ambas operaciones consisten en una mera búsqueda binaria, que en el caso de la inserción al encontrar un nodo vacío (*None*) inserta un nuevo nodo en ese espacio. Las operaciones se implementaron de forma iterativa en vez de recursiva para evitar el desborde de la pila al insertar secuencias ordenadas extensas, ya que Python no provee optimización de llamada por la cola.

2.3. ABB Aleatorizado

En cada ABB aleatorizado T , debe cumplirse que para todo sub-árbol T' cada nodo tiene probabilidad $\frac{1}{|T'|}$ de ser la raíz, donde $|T'|$ es la cantidad de nodos de T' . Esta propiedad se cumple cuando al insertar un nodo en un sub-árbol T' , este nodo tiene probabilidad $\frac{1}{|T'|+1}$ de convertirse en la raíz de T' . Esto puede lograrse generando un número n pseudoaleatorio distribuido uniformemente entre 0 y $|T'|$, y luego si $n = |T'|$ insertamos al nodo como raíz de T' . Para calcular $|T'|$ en cada T' fue necesario agregar el campo *children* a la clase *Node*, el cual almacena la cantidad de descendientes de este nodo.

Para que un nodo se convierta en raíz de un árbol manteniendo la consistencia del mismo lo que se hace es rotarlo desde su posición de inserción hasta que quede en la raíz. A continuación se muestra el algoritmo de inserción, dividido entre la inserción normal y la inserción en la raíz.

insert(k):

- $\text{self.root} := \text{insert_rec}(\text{self.root}, k)$

insert_rec(node, k):

- Si $node = None$, retorna un nuevo nodo de llave k .
- Si $random(0, node.children + 1) = node.children$
 - Retorna $insert_at_root(node, k)$
- Si $k < node.key$, $node.left := insert_rec(node.left, k)$
- En caso contrario, $node.right := insert_rec(node.right, k)$
- $node.children := node.children + 1$
- Retorna $node$

insert_at_root(node, k):

- Si $node = None$, retorna un nuevo nodo de llave k .
- Si $k < node.key$:
 - $node.left := insert_at_root(node.left, k)$
 - $node = rotate_right(node)$
- En caso contrario:
 - $node.right := insert_at_root(node.right, k)$
 - $node = rotate_left(node)$
- Retorna $node$

Las rotaciones de nodo se implementaron según el esquema de la figura 2. Estas además actualizan el campo *children* del nodo rotado.

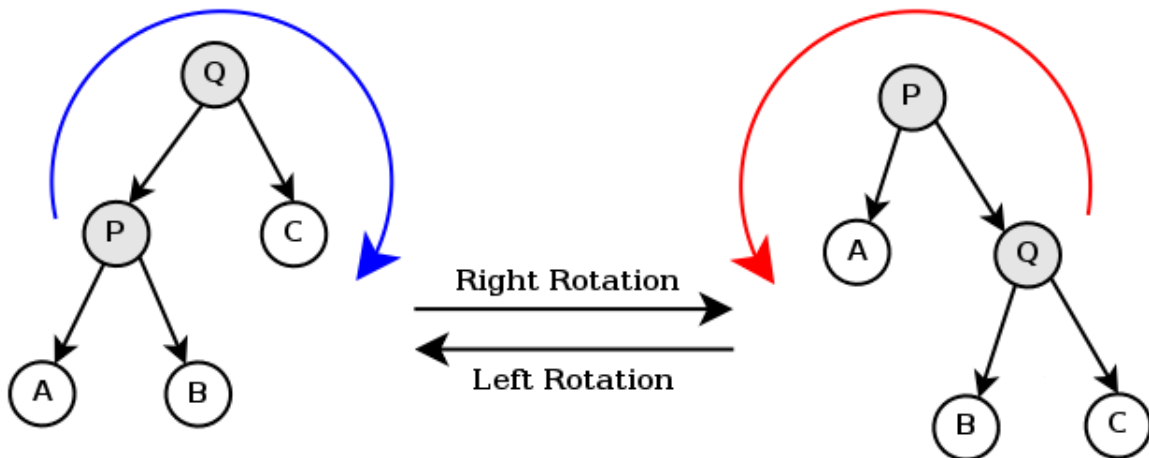


Figura 2: Operaciones de rotación de nodos, hacia la derecha e izquierda.

3. Experimentos

Se llevaron a cabo los siguientes experimentos:

1. Generar una secuencia ordenada de n números; donde $n \in \{10^4, 2 \times 10^4, 5 \times 10^4\}$ y los números corresponden a un subconjunto de enteros entre $[0, 10^5]$.
2. Repetir hasta que, para la secuencia ingresada, se obtenga un comportamiento similar entre los ABB normales y los ABB aleatorizados (se consideró una diferencia de 10 %, es decir, detenerse mientras la cantidad de comparaciones del ABB normal sea mayor en un 10 % a la de un ABB aleatorizado):
 - 2.1 Almacenar en cada estructura de datos una secuencia de n números.
 - 2.1.1 En los ABB normales se aplican k swaps antes de realizar la secuencia de inserciones; donde $k = 0,005 \times n$. En cada swap se seleccionan aleatoriamente 2 elementos de la secuencia de entrada y se intercambian.
 - 2.2 Buscar m valores en la estructura de datos; donde $m = 0,5 \times n$ y de los cuales 25 % corresponden a búsquedas infructuosas.
 - 2.3 Utilizar la secuencia obtenida después de aplicar los k swaps en la siguiente iteración.

Los resultados se presentan en las secciones siguientes.

3.1. Tiempos de inserción y búsqueda

Se calcularon los tiempos de inserción y búsqueda promedio en función de n , a partir de diversas repeticiones del experimento explicado en la sección 3. Estos tiempos se midieron como comparaciones de llaves. A continuación se muestran los gráficos que muestran el comportamiento de las operaciones, y las curvas de ajuste que se encontraron.

3.1.1. Inserción en ABB aleatorizado

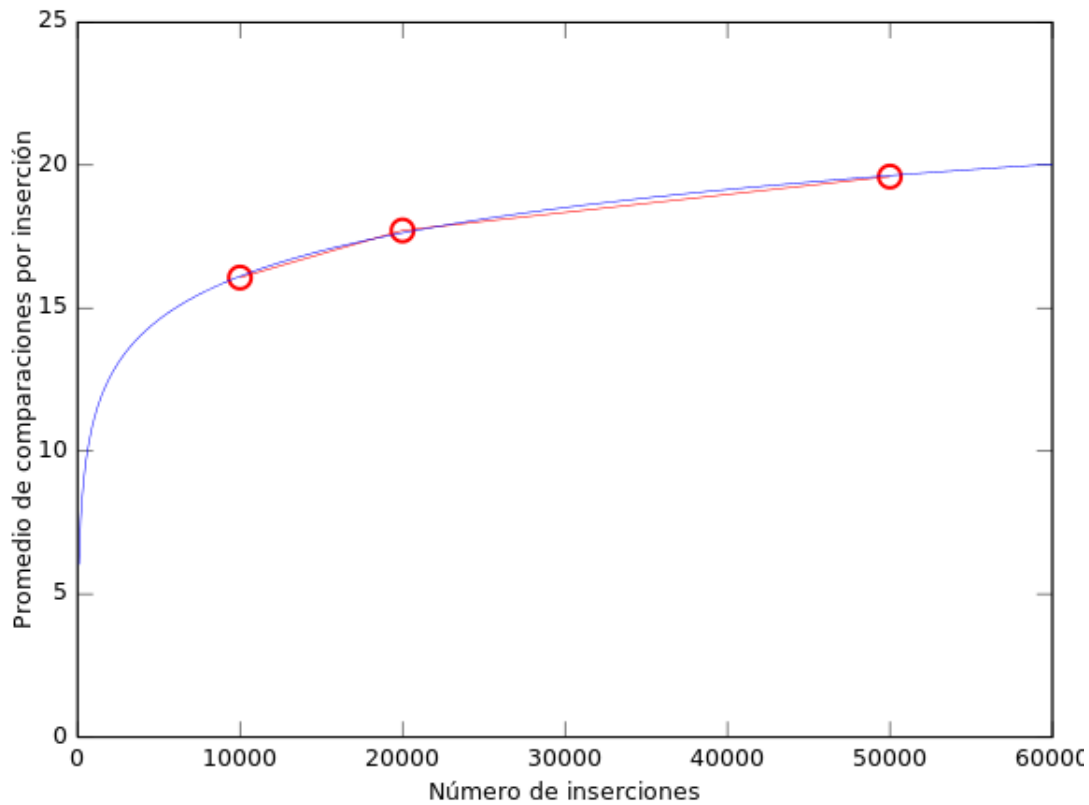


Figura 3: Inserción en ABB aleatorizado.

Modelo de ajuste: $2,1886 \times \log(n) - 4,0455$

3.1.2. Inserción en Skip Lists

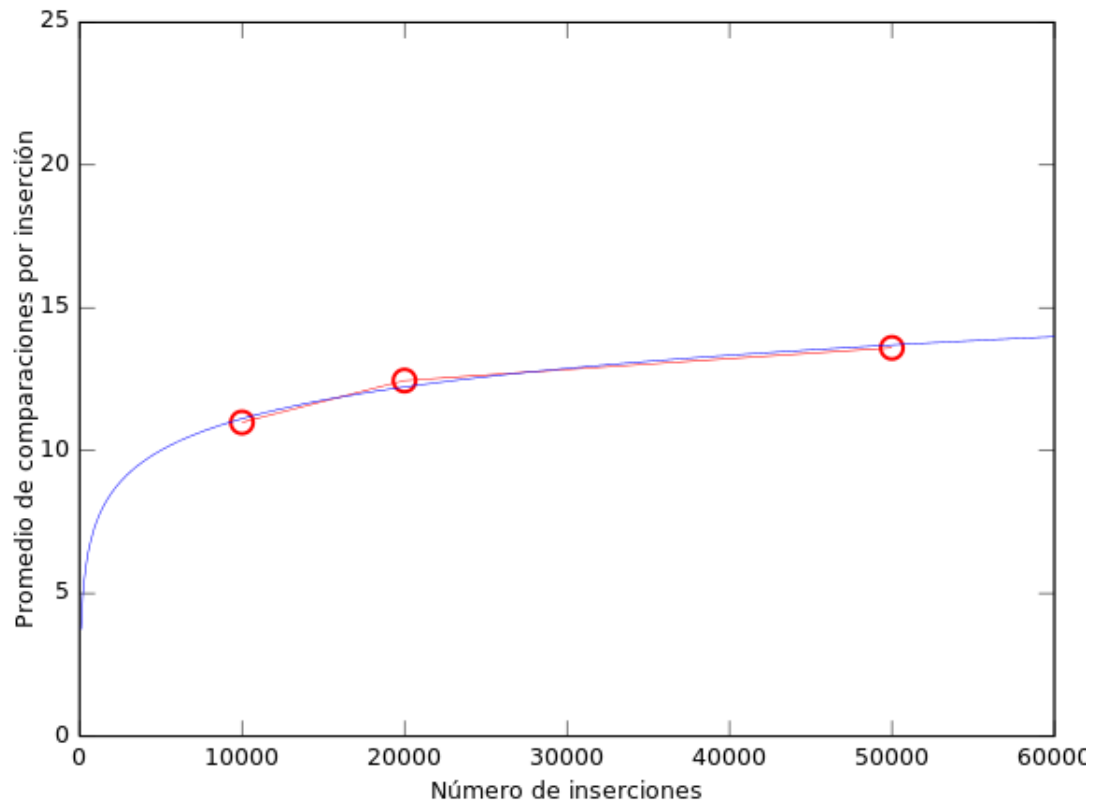


Figura 4: Inserción en Skip List.

Modelo de ajuste: $1,6022 \times \log(n) - 3,6407$

3.1.3. Búsqueda en ABB aleatorizado

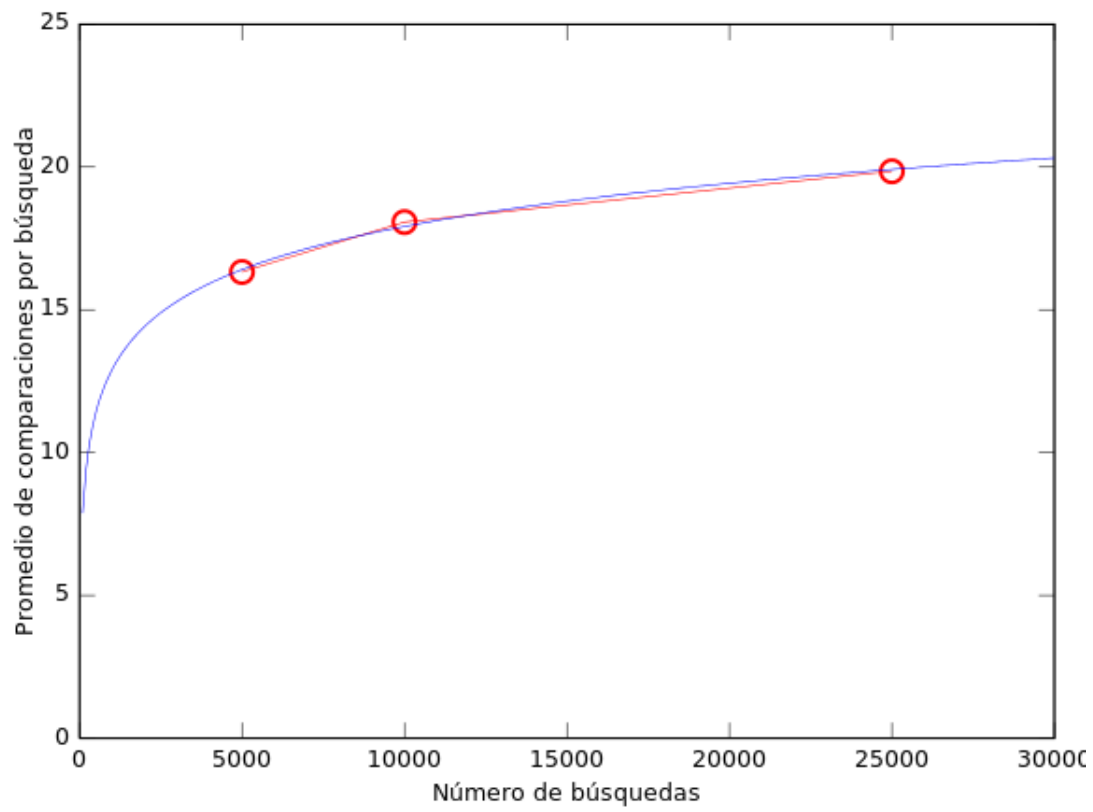


Figura 5: Búsqueda en ABB aleatorizado.

Modelo de ajuste: $2,1799 \times \log(n) - 2,1612$

3.1.4. Búsqueda en Skip Lists

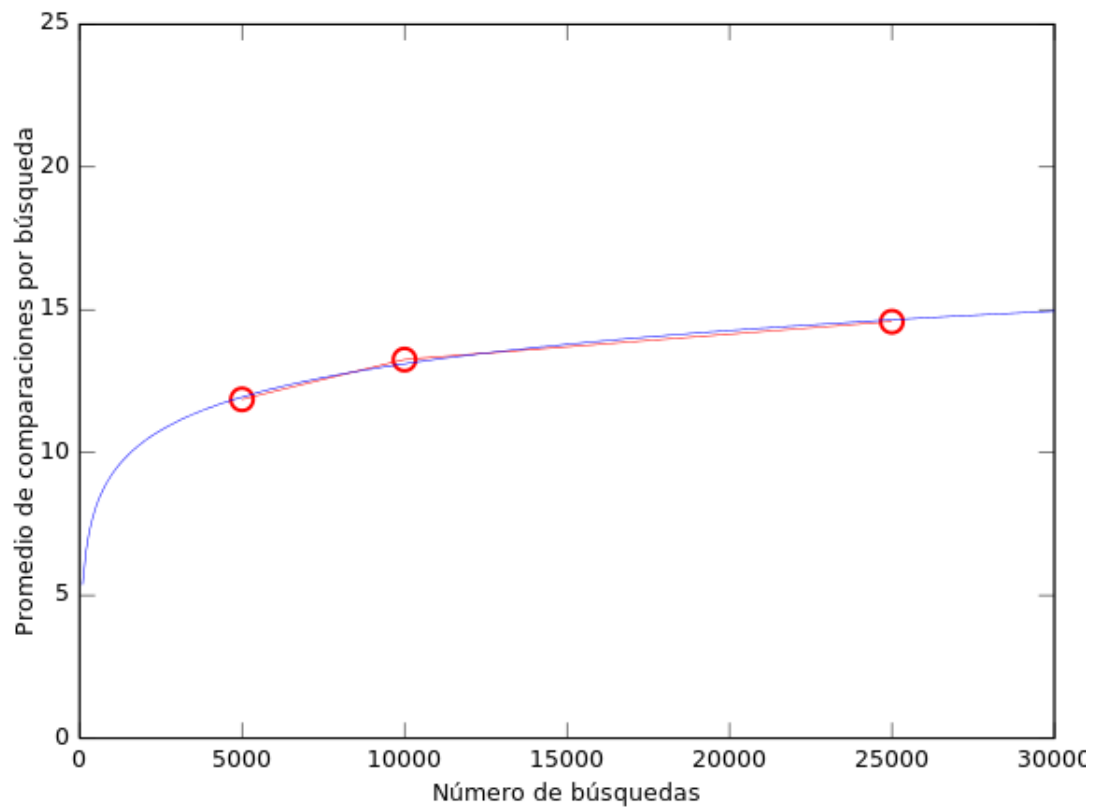


Figura 6: Búsqueda en Skip List.

Modelo de ajuste: $1,6797 \times \log(n) - 2,3632$

3.2. Comparación entre ABB y ABB aleatorizado

El diseño del experimento garantiza una convergencia en cuanto a la cantidad de comparaciones entre el ABB y el ABB aleatorizado (cada iteración disminuye las comparaciones de un ABB normal y se acerca más a un ABB aleatorizado).

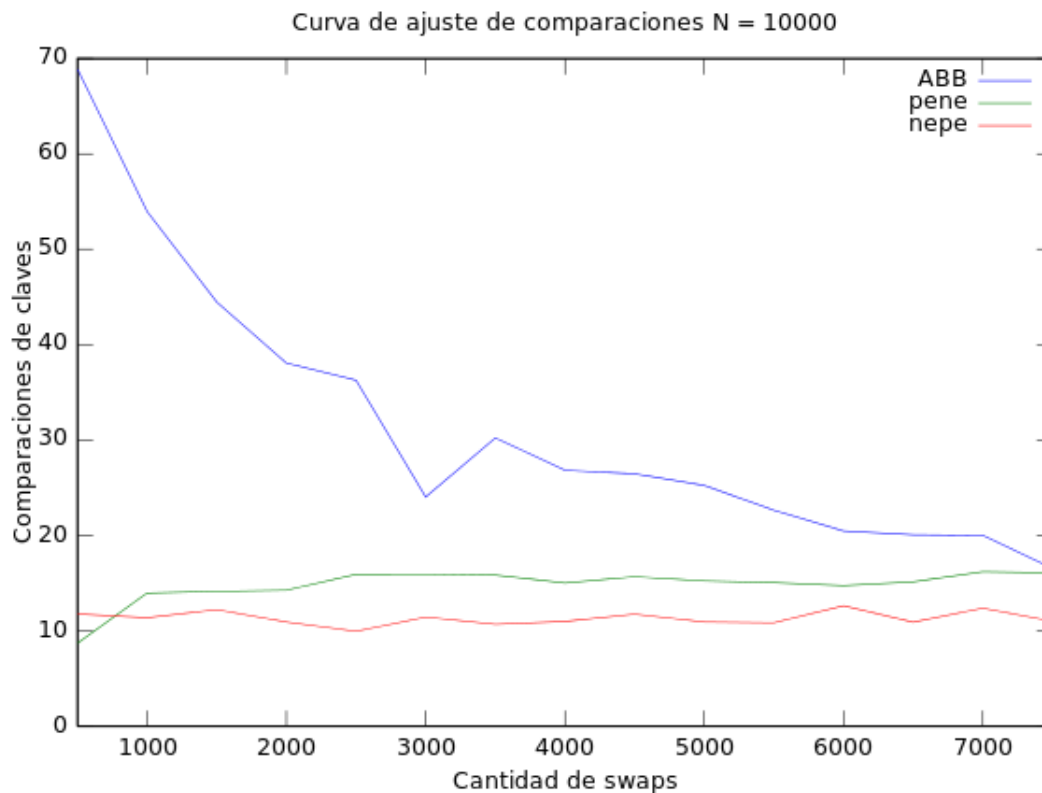


Figura 7: Comparación de claves vs cantidad de swaps sobre la entrada..

En la figura 8 se puede apreciar, en el caso de $n = 10000$ la convergencia en cuanto a comparaciones del ABB (y además la pseudo independencia a la entrada tanto del ABB aleatorizado y de la SkipList).

Ahora, los resultados fueron los siguientes:

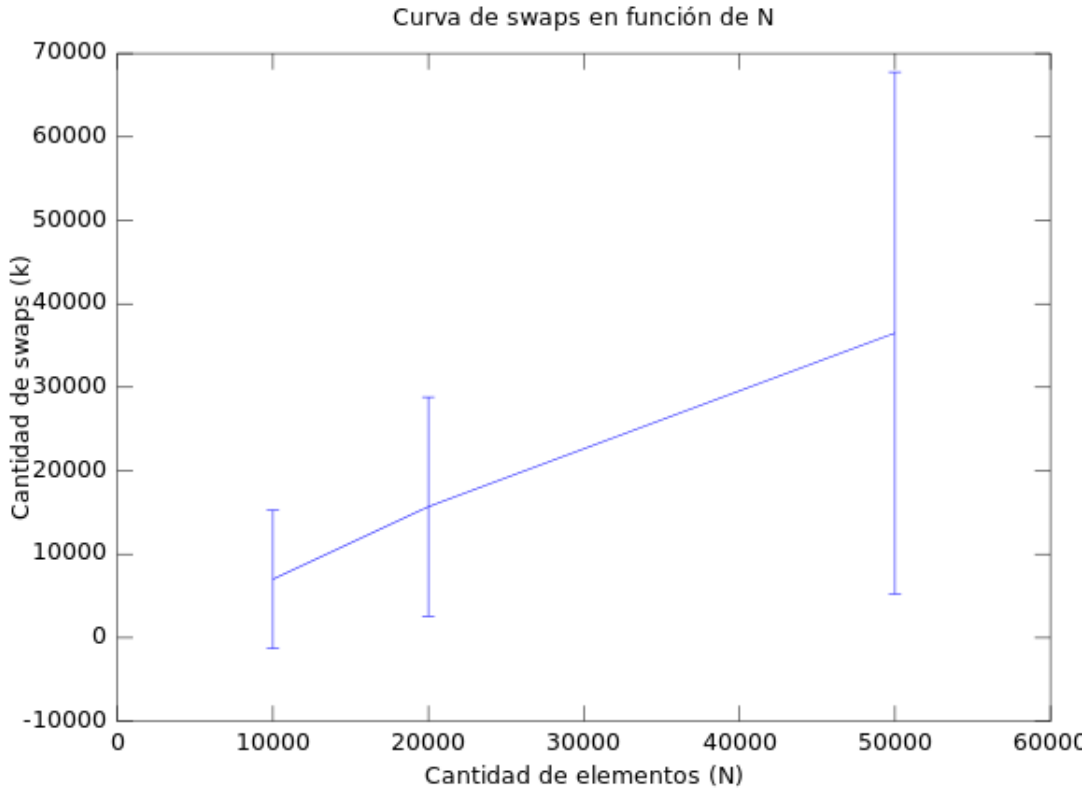


Figura 8: Swaps vs cantidad de elementos.

Los ABB normales mostraron ser vulnerables a la entrada y alcanzar una cercanía en cuanto a comportamiento a un ABB aleatorizado después de una cantidad cercana a $6/10 * n$ en promedio.

4. Conclusiones

Los ABB son absolutamente vulnerables al orden en la entrada, y por ende, su convergencia en cuanto a comportamiento se alcanza luego de que el arreglo está completamente desordenado en promedio (la cantidad de swaps necesario para alcanzar este comportamiento es cercano a $7/10 * n$). Una vez que se alcanza este parecido en comportamiento con los ABB aleatorizados, se puede concluir que se comportan de igual forma, y además, que los ABB aleatorizados son inmunes a la entrada.

Sin embargo, de los resultados obtenidos en la sección 3.1 podemos observar que las Skip List se comportan mejor que los ABB, lo que se deduce a partir de que las constantes que multiplican al logaritmo en los modelos de ajuste son menores tanto en el caso de la inserción como la búsqueda. Esto se debe a que su aleatorización tiene una independencia estructural muy grande. Mientras los ABB aleatorizados deben efectuar rotaciones en base a un árbol no perfectamente balanceado las SkipList agregan atajos con cierta probabilidad, lo que, visto a nivel macro, es equivalente a rehacer completamente el árbol y mantener además el invariante de balanceo de hijos. Una SkipList se comporta en promedio de igual forma que un árbol de búsqueda binario

perfectamente balanceado lo que implica que su altura es $\lceil \log_2 n \rceil$ con alta probabilidad.