



Representaciones Sucintas de Árboles, con Aplicación en consultas Xpath

Diego Arroyuelo

Yahoo! Research Latin America

darroyue@yahoo-inc.com

A photograph of a desert landscape with sand dunes and footprints. The sand is light-colored and has a wavy, rippled texture. Several dark footprints are visible, leading from the foreground towards the background. The sky is a clear, pale blue.

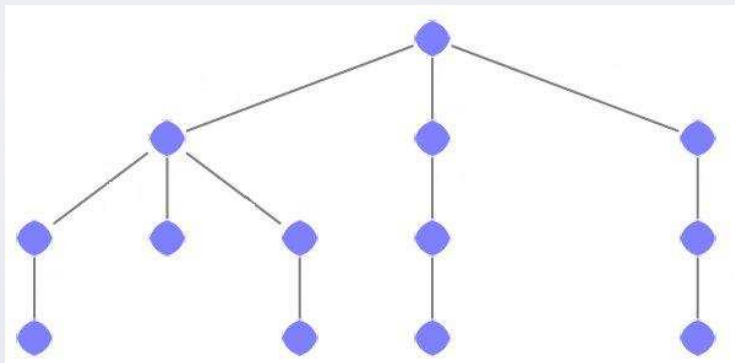
Introducción y Motivación

Introducción

- Los árboles son una estructura de datos básica en CS
- La representación clásica requiere demasiado espacio
 - $O(n \log n)$ bits para representar un árbol de n nodos
 - Espacio extra para soportar operaciones adicionales
- Problema: demanda creciente de almacenamiento de información
 - Árboles con millones de nodos y necesidad de soportar varias operaciones
- Estructuras de datos sucintas...

Introducción

- Representaciones sucintas de árboles:
 - $2n + o(n)$ bits
 - Soportan un gran número de operaciones en tiempo $O(1)$
- Por ejemplo, representación de paréntesis balanceados



→ (((()))(())(()))(())(())(())(())

Introducción

- Operaciones básicas sobre árboles

pre-rank(i) / post-rank(i)
pre-select(i) / post-select(i)
isleaf(i)
isancestor(i, j)
depth(i)
parent(i)
first-child(i) / last-child(i)
next-sibling(i) / prev-sibling(i)
subtree-size(i)
level-ancestor(i, d)

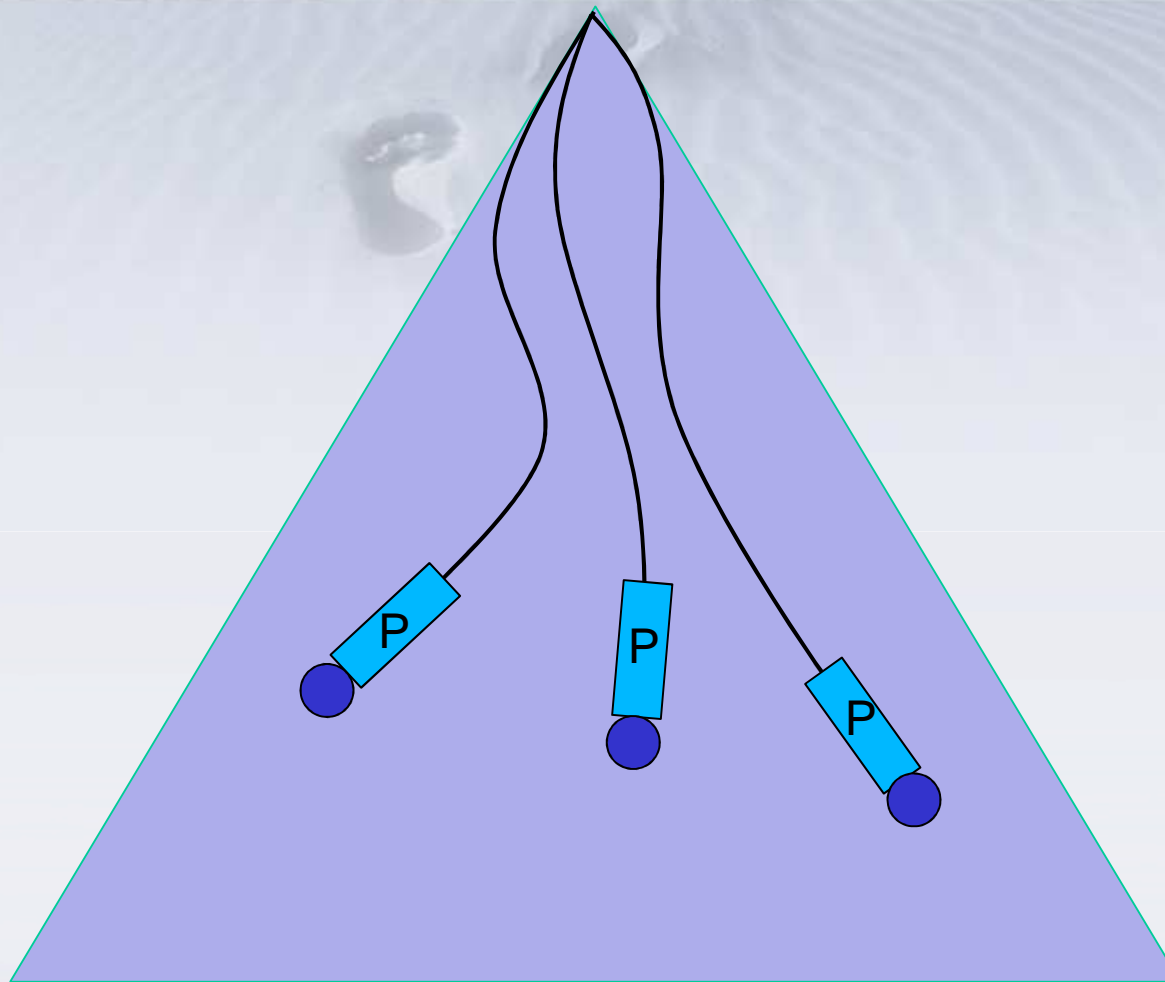
lca(i, j)
deepest-node(i)
degree(i)
child(i, q)
child-rank(i)
in-rank(i)
in-select(i)
leaf-rank(i)
leaf-select(i)

Todas soportadas en tiempo $O(1)$

Introducción

- Aplicación típica: búsqueda por prefijos (búsqueda en texto, etc.)
- Otras aplicaciones necesitan operaciones más complicadas sobre árboles
- Subpath(P): encontrar las ocurrencias del patrón $P[1..m]$ en los caminos del árbol
 - Tree pattern matching [**Kosaraju – FOCS'89**]
 - Búsqueda en texto con LZ-índices [**Arroyuelo et al. – CPM'06**]
 - Consultas Xpath [**Ferragina et al. – JACM 2009**]

Introducción

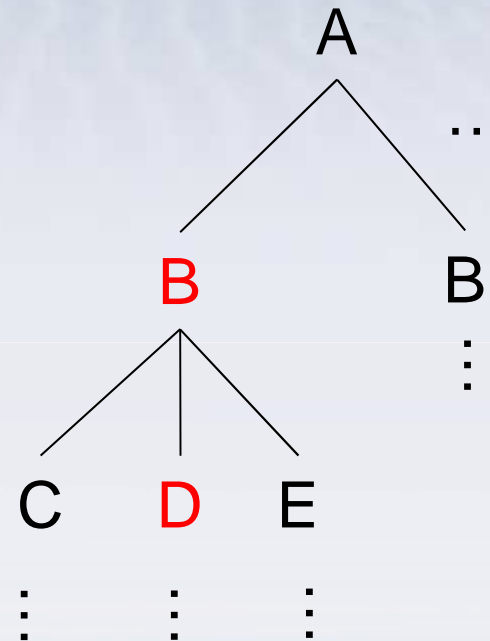


Estamos interesados en representaciones que usen poco espacio y soporten operaciones básicas y subpath (de manera intercalada)

Introducción

Ejemplo de búsqueda tipo XPath

```
<A>
  <B>
    <C> ... </C>
    <D> ... </D>
    <E> ... </E>
  </B>
  <B>
    ...
  </B>
  ...
</A>
```



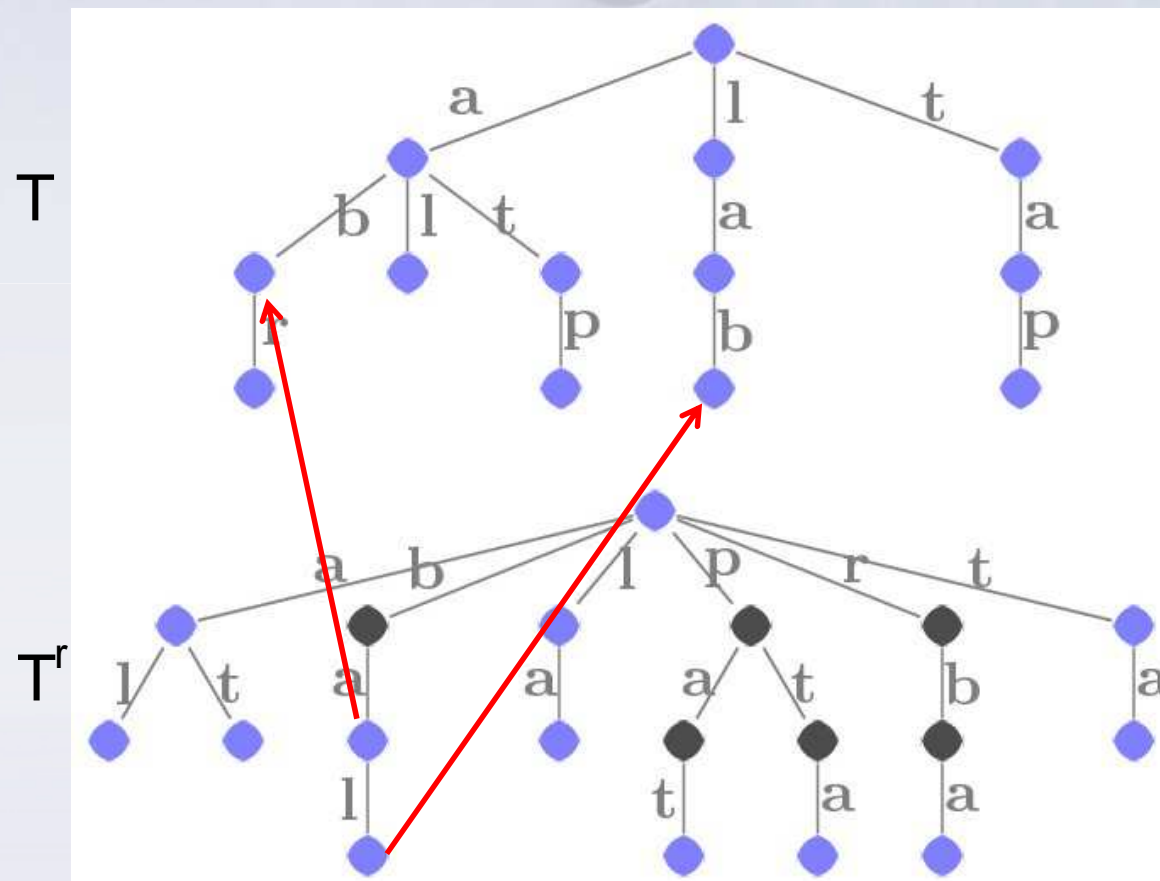
Consulta: //BD

A photograph of a desert landscape with sand dunes. The sand is light-colored and has fine, wavy ripples. In the center, there are several dark, irregular footprints or tracks, suggesting someone has walked through the sand. The sky is a clear, pale blue.

Trabajo Relacionado

Trabajo relacionado

- Kosaraju [FOCS'89] propone usar árbol de sufijos de los caminos del árbol.



Respuesta a una subpath es un subárbol en el árbol reverso

Requiere mucho espacio (punteros más conexión entre árboles)

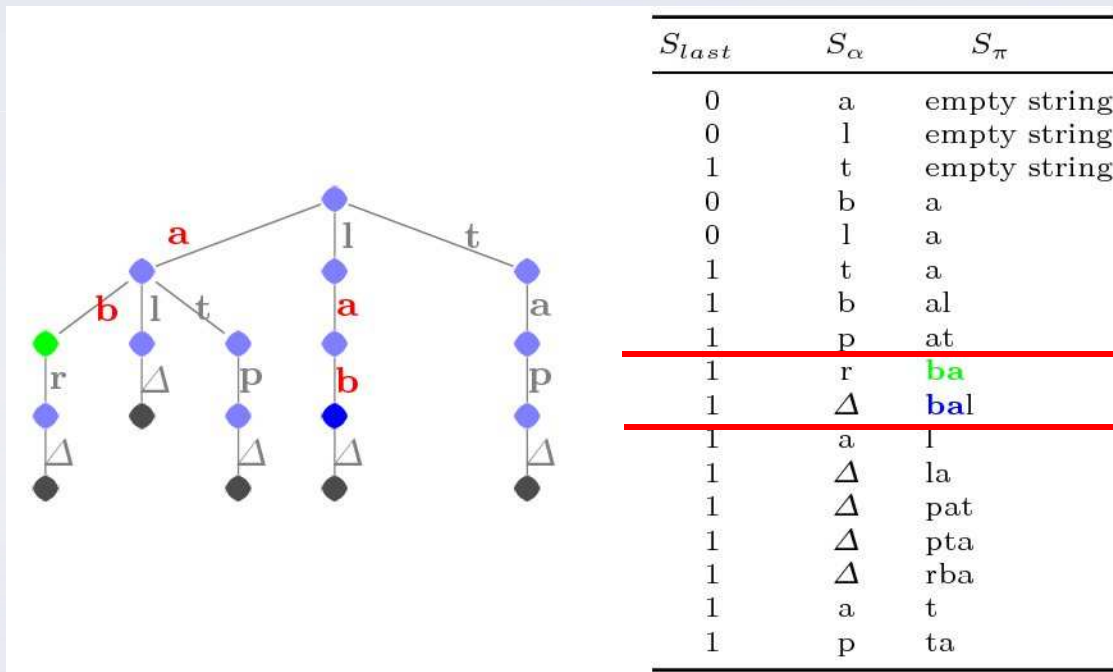
Puede representarse con menos espacio...

Trabajo relacionado

- La solución de Kosaraju puede representarse con menos espacio:
 - Árboles con representación sucinta
 - Conexión entre árboles con la idea de Arroyuelo et al. [CPM'06] **($\epsilon n \log n$ bits - tiempo $O(1/\epsilon)$)**
- Espacio: $6n + 3n \log \sigma + 2\epsilon n \log n + o(n \log \sigma)$ bits
- Tiempo:
 - $O(1)$ para operaciones básicas
 - $O(m/\epsilon)$ para subpath
 - $O(1/\epsilon)$ para operaciones básicas después de una subpath

Trabajo relacionado

- Ferragina et al. [JACM, 2009] definen la transformada xbw
 - Extensión a árboles de la transformada de Burrows-Wheeler



Respuesta a una subpath es un intervalo de la transformada

Trabajo relacionado

- La xbwt no provee soporte natural para muchas de las operaciones básicas (pre-rank, depth, subtree-size, etc.):
 - Es una permutación de los nodos del árbol
- Alternativa para soportar todas las operaciones:
 - Transformada xbw del árbol
 - Representación sucinta del árbol original
 - Conexión entre representaciones (**$\epsilon n \log n$ bits – tiempo $O(1/\epsilon)$**)
- Espacio: $6n + 5n \log \sigma + 2\epsilon n \log n + o(n \log \sigma)$ bits
- Tiempo:
 - $O(1)$ para operaciones básicas,
 - $O(m)$ para subpath
 - $O(1/\epsilon)$ para operaciones básicas después de una subpath

Trabajo relacionado

- Tenemos dos soluciones para consultas subpath

1. **Árbol de sufijos de los caminos del árbol**

- Espacio: $6n + 3n \log \sigma + 2\epsilon n \log n + o(n \log \sigma)$ bits

- Tiempo:

- $O(1)$ para operaciones básicas
- $O(m/\epsilon)$ para subpath
- $O(1/\epsilon)$ para operaciones básicas después de una subpath

2. **Transformada xbw**

- Espacio: $6n + 5n \log \sigma + 2\epsilon n \log n + o(n \log \sigma)$ bits

- Tiempo:

- $O(1)$ para operaciones básicas,
- $O(m)$ para subpath
- $O(1/\epsilon)$ para operaciones básicas después de una subpath

A photograph of a desert landscape with sand dunes. The dunes are covered in fine, ripples of sand. In the center, there are several footprints, suggesting someone has walked through the sand. The sky is a clear, pale blue.

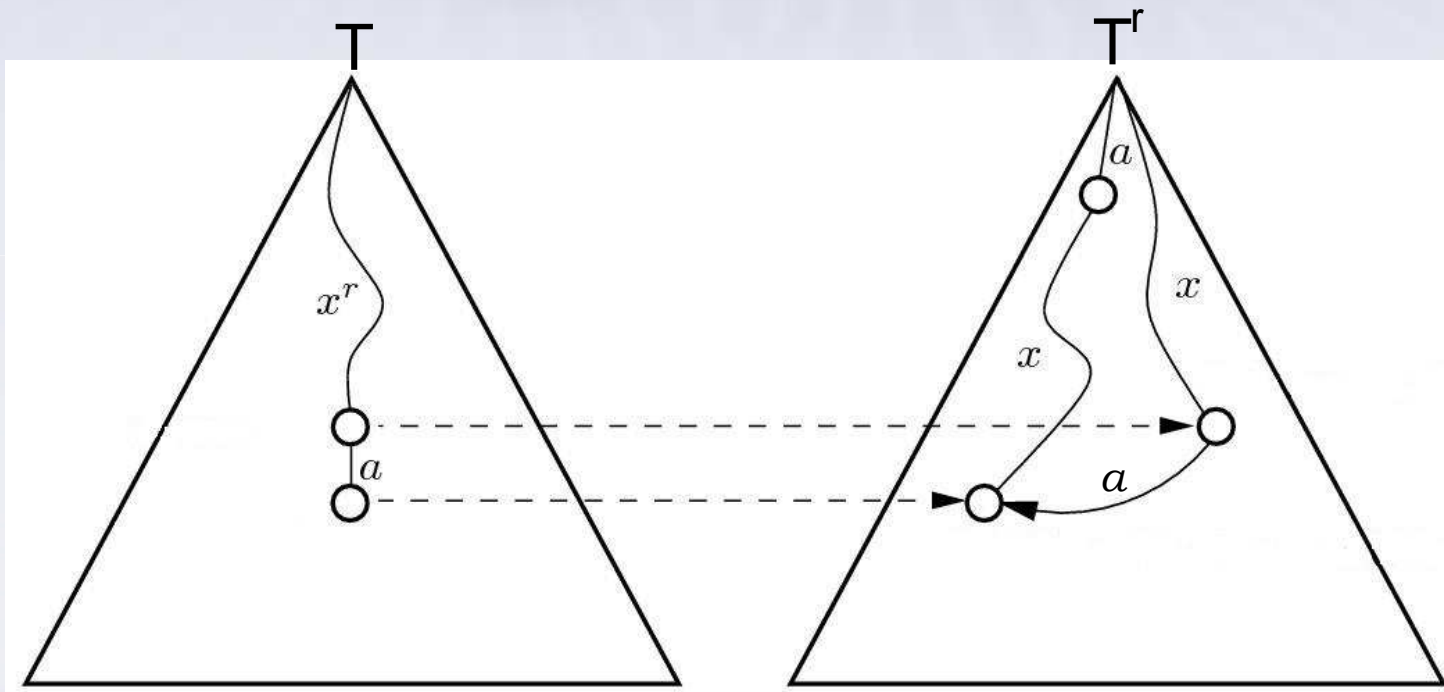
Una Nueva Representación

Una nueva representación

- Ninguna de esas representaciones soporta la intercalación de operaciones de forma natural
- En busca de una representación que soporte subpaths sobre la estructura original del árbol...
- La idea es simular la estructura del árbol reverso sobre los nodos del árbol original
- Estructura de Bi-trees [Stoye, 1996]:
 - Estructuras distintas sobre el mismo conjunto de nodos

Una nueva representación

- Cómo simular la estructura del árbol reverso sobre el árbol original?



Seguir un Weiner link en el reverso es equivalente a ir al hijo en el árbol original

Una nueva representación

- Los nodos ficticios se agregan en la fringe del árbol
- Resultado:
 - La estructura original del árbol se mantiene
- Eso simplifica soportar las operaciones básicas sobre el árbol
- Para subpaths se navega sobre los Weiner links usando P^r
 - La profundidad de cada nodo indica el símbolo a usar en cada nodo
 - El chequeo del final se hace siguiendo el camino hacia la raíz

Una nueva representación

- Espacio: $8n + 4n \log \sigma + 2n \log n + o(n \log \sigma)$ bits
- Tiempo:
 - $O(1)$ para operaciones básicas
 - $O(m)$ para subpath
 - $O(1)$ para operaciones básicas después de una subpath
- En la práctica, en algunos casos el espacio puede ser cercano a:
 - $6n + 2n \log \sigma + n \log n + o(n \log \sigma)$ bits

A photograph of a desert landscape featuring sand dunes with fine, rhythmic ripples. A series of footprints is visible, leading from the foreground towards the background. The sky is a clear, pale blue.

Conclusión

Conclusión

- En búsqueda de una representación de árboles que soporte conjunto completo de operaciones + subpaths
- La representación de Bi-trees es interesante:
 - Menor espacio en la práctica
 - Una única estructura puede beneficiar la localidad de acceso a ella
 - Puede soportar subpaths a partir de cualquier nodo
- Problema abierto: lograr que la representación use $O(n \log \sigma)$ bits de espacio [Ferragina et al., JACM 2009]
 - Adaptar la representación de Arroyuelo et al. [TR/DCC, 2009] para representar Weiner links con $O(n \log \sigma)$ bits

A photograph of a desert landscape featuring sand dunes with fine, wavy ripples. A trail of footprints is visible, leading from the foreground towards the background. The sky is a clear, pale blue.

¿Preguntas?



Gracias!