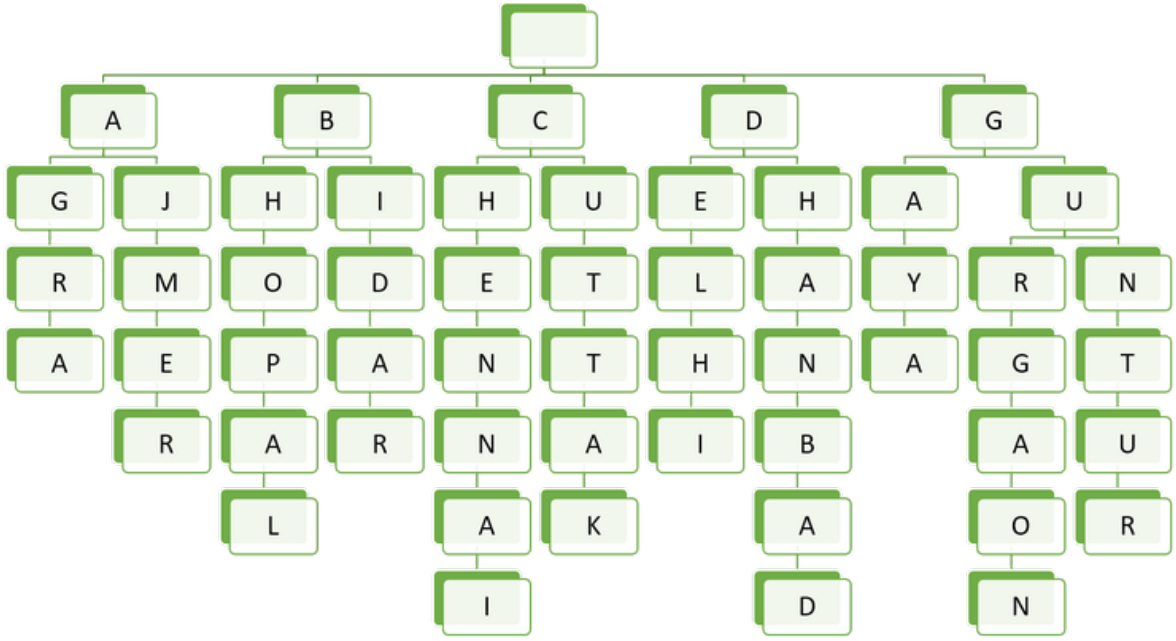
MAC0323 Algoritmos e Estruturas de Dados II

Edição 2020 – 2

AULA 17

Tries (árvores digitais)



Fonte: `Building an autocomplete system using Trie`

Referências: Tries (árvores digitais) (PF); Tries (S&W); slides (S&W); Vídeo (S&W); TAOCP, vol 3, cap. 6.3

R-way tries

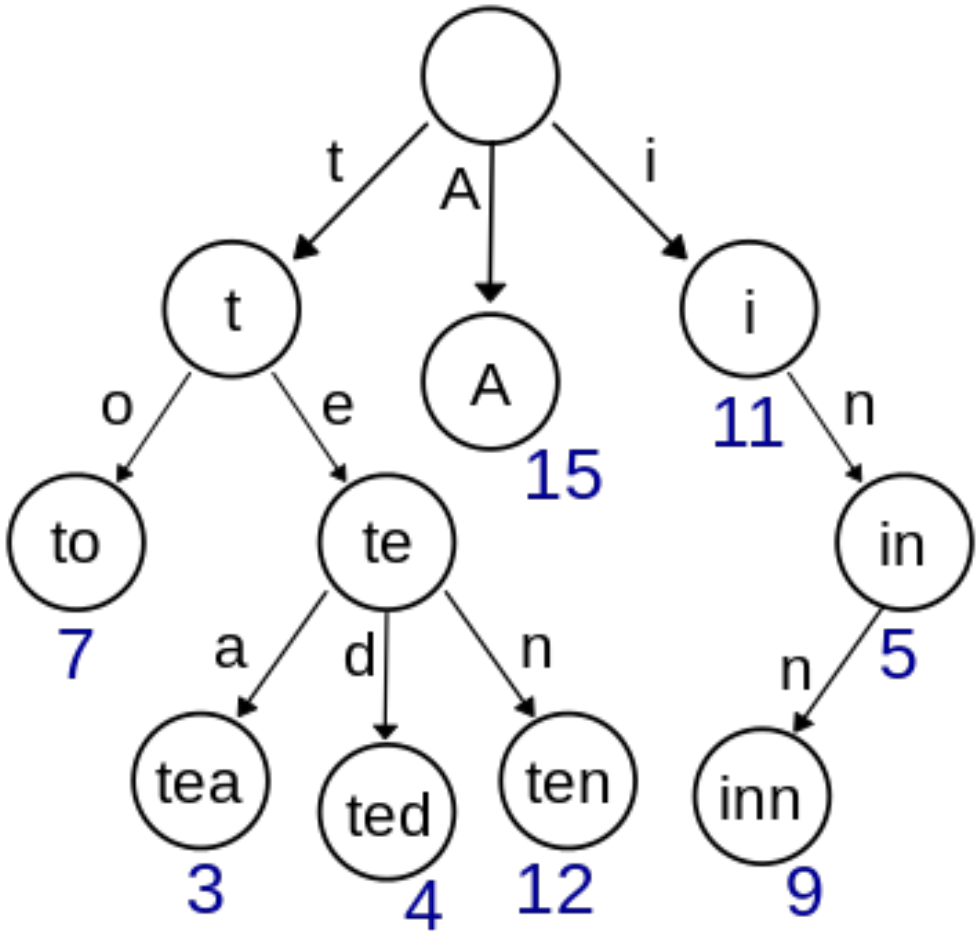
Uma **trie** (=*R-way trie*) é um tipo de árvore usado para implementar STs de strings sobre um alfabeto com R símbolos.

Tries também são conhecidas como árvores digitais e como árvores de prefixos.

Com Tries, em vez do método de busca ser baseado em comparações entre chaves, é utilizada a representação das chaves como caracteres de um alfabeto.

Considere, por exemplo,
a busca de uma palavra no dicionário:
a primeira letra indica as páginas que devemos olhar;
a segunda letra restringe o espaço de busca; . . .

Ilustração



Fonte: Wikipedia

Métodos específicos

A `API` de uma `trie` inclui, além das rotinas usuais como `put()`, `get()`, `delete()`, ...
três métodos específicos:

- `keysWithPrefix(char *s)`:
  todas as chaves que têm prefixo `s`;
- `keysThatMatch(char *s)`:
  todas as chaves que *casam* com `s`
  quando '`.`' é usado como **curinga**;
- `longestPrefixOf(char *s)`:
  a chave mais longa que é prefixo de `s`.

Métodos específicos

Exemplos para o conjunto de chaves

`she sells sea shells by the sea shore`:

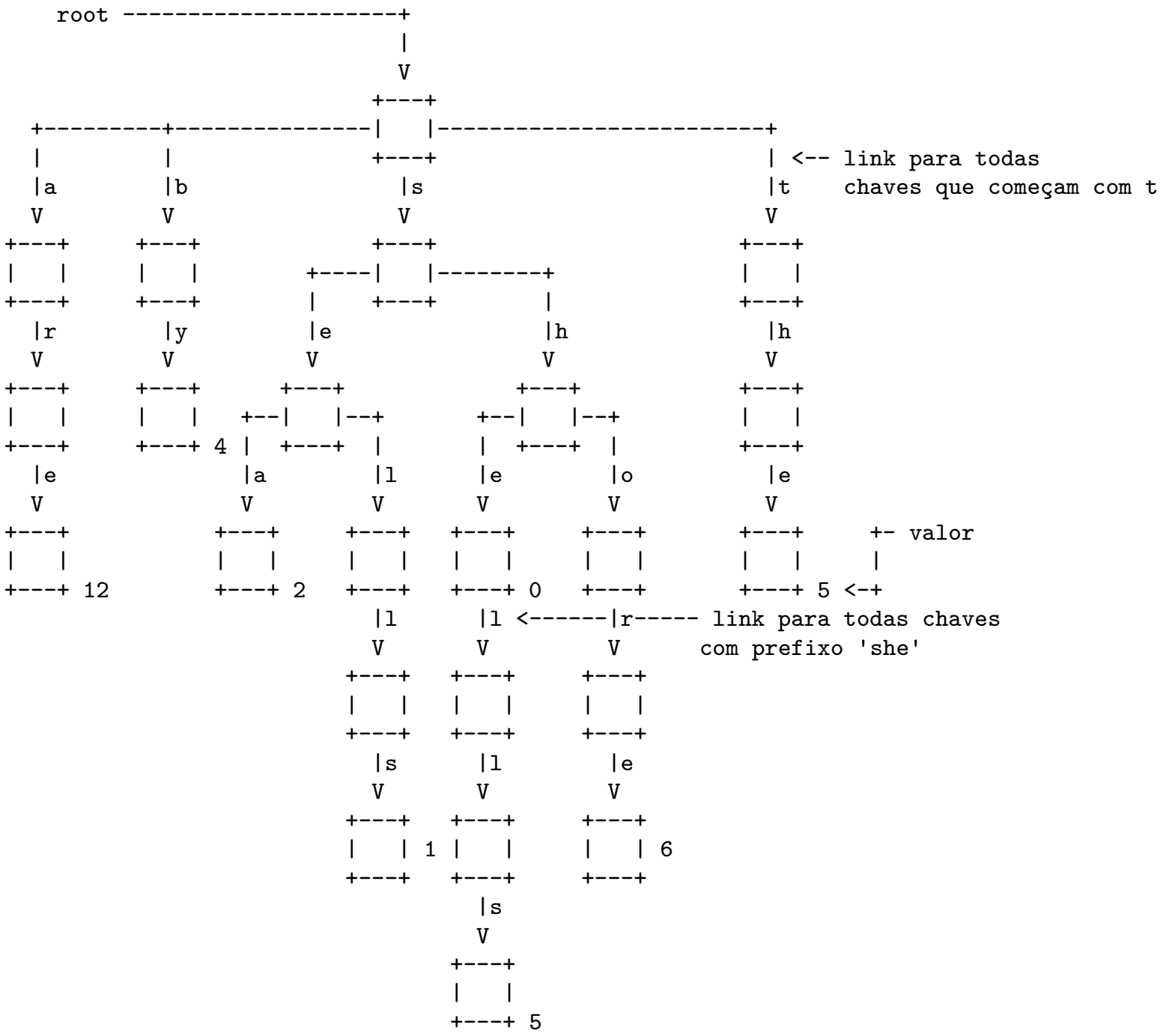
```
keysWithPrefix("she") devolve "she" e "shells"
keysWithPrefix("se") devolve "sells" e "sea"
keysThatMatch(".he") devolve "she" e "the"
keysThatMatch("s..") devolve "she" e "sea"
longestPrefixOf("shell") devolve "she"
longestPrefixOf("shellsort") devolve "shells"
```

Outra ilustração

Para os pares `key`-`val`

`key`	`val`
are	12
by	4
sea	2
sells	1
she	0
shells	3
the	5
shore	6

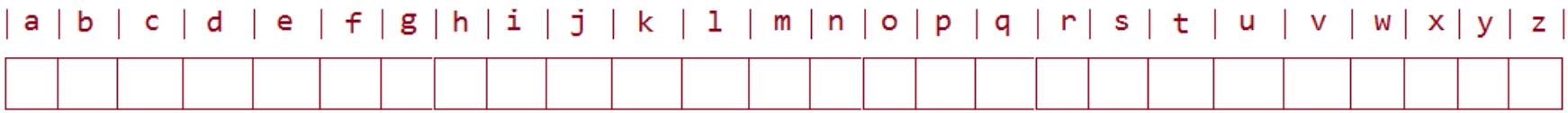
temos a `trie` a seguir.


root
<-- link para todas
chaves que começam com t
4
12
2
0
5 <-+
+- valor
link para todas chaves
com prefixo 'she'
1
6
5

Estrutura do nó de uma `trie`

Os links correspondem a caracteres e não a chaves.
Tries são compostas por nós do tipo `Node`.

```
static struct Node {
   Value val;
   Node next[R];
}
```

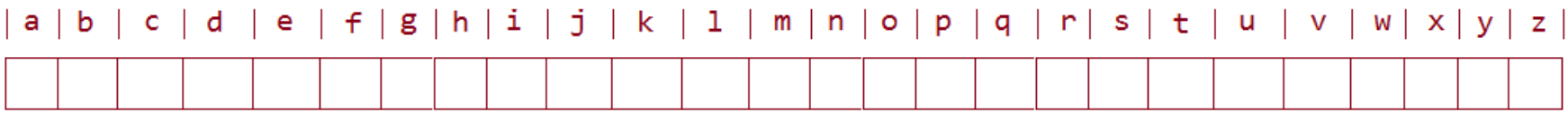


Fonte: `JavaByPatel`

Muitos dos `R` ponteiros podem ser `NULL`.

Ilustração

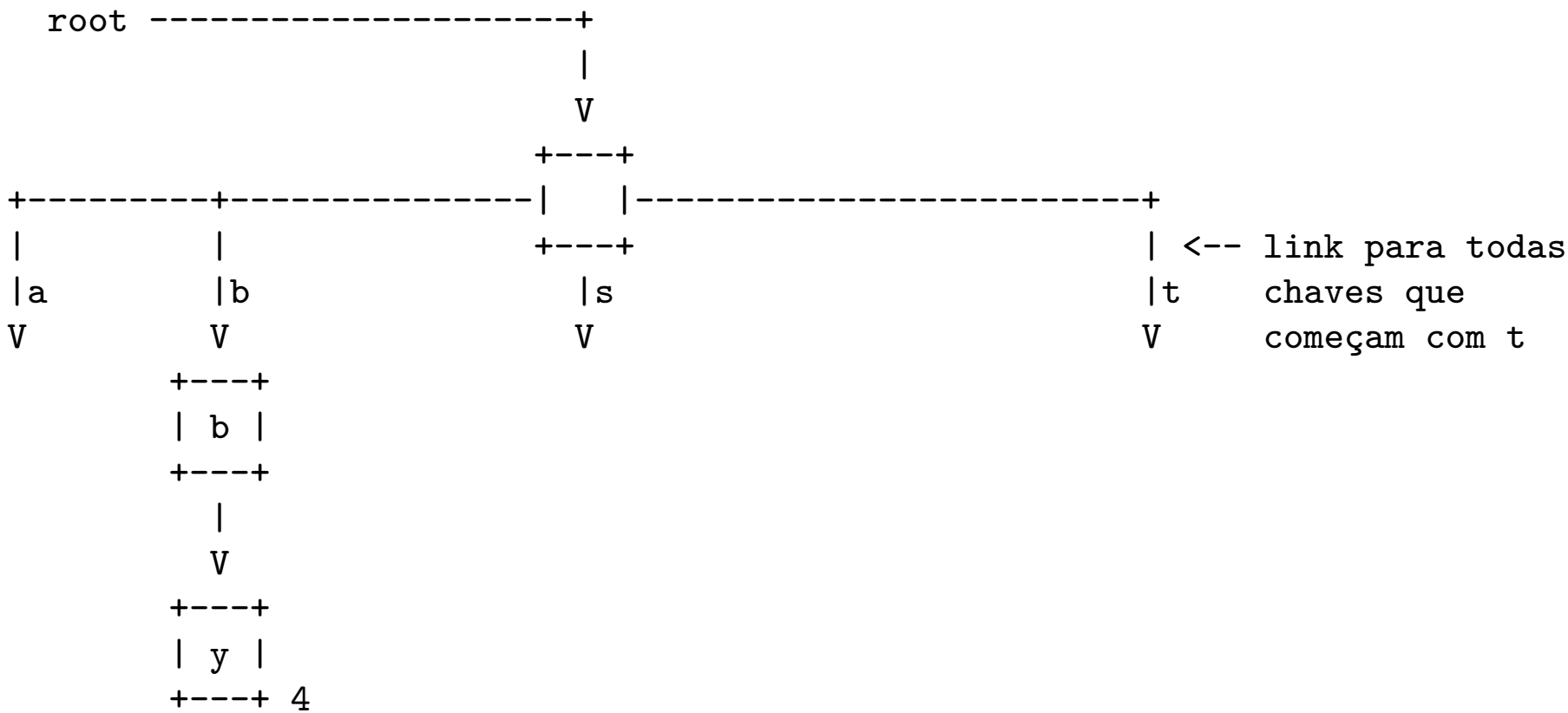
Se a trie é para o alfabeto `'a', 'b',...,'z'` temos



Fonte: `JavaByPatel`

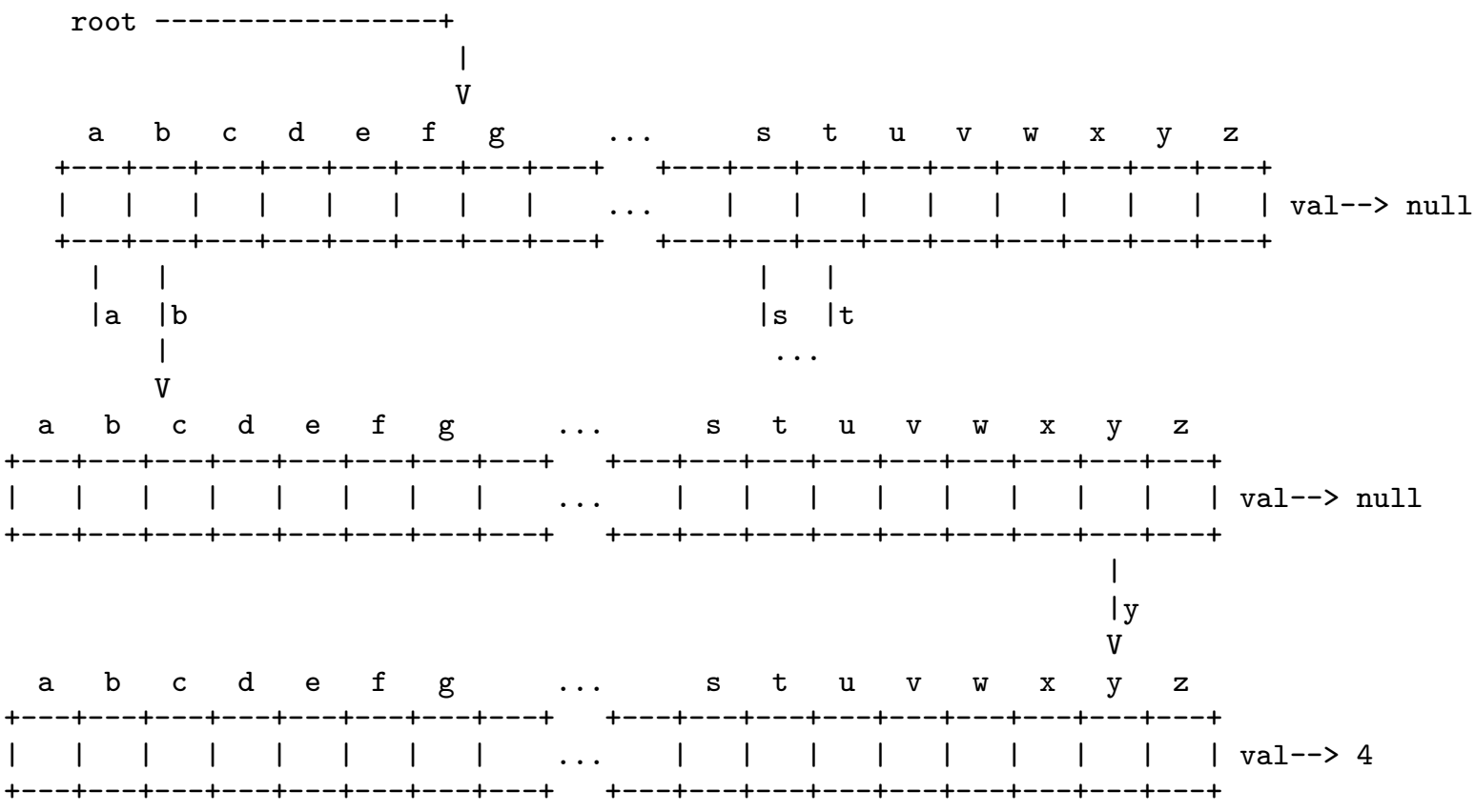
Ilustração

Se a trie é para o alfabeto `'a', 'b',...,'z'` temos



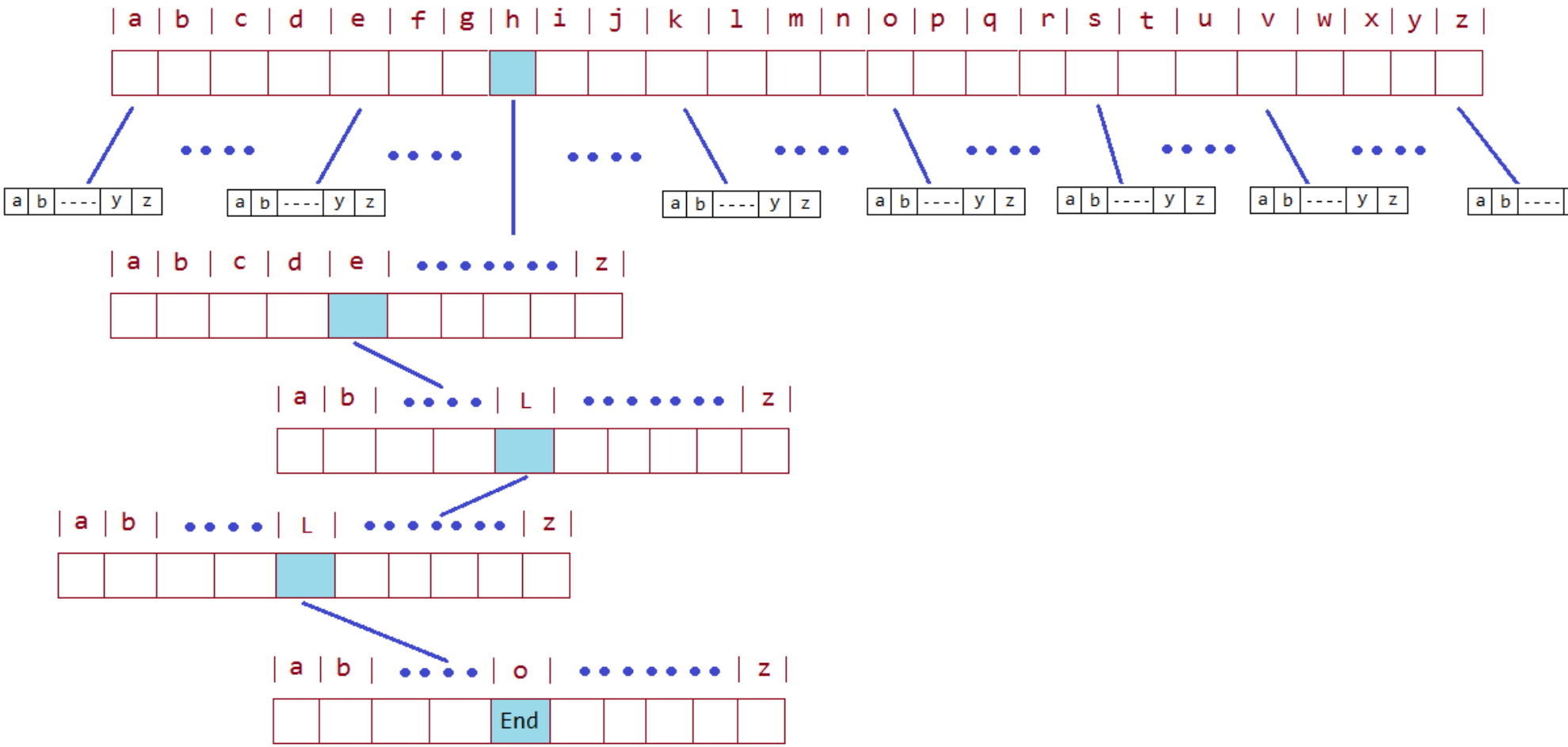
Outra ilustração

Se a trie é para o alfabeto `'a', 'b',...,'z'` temos



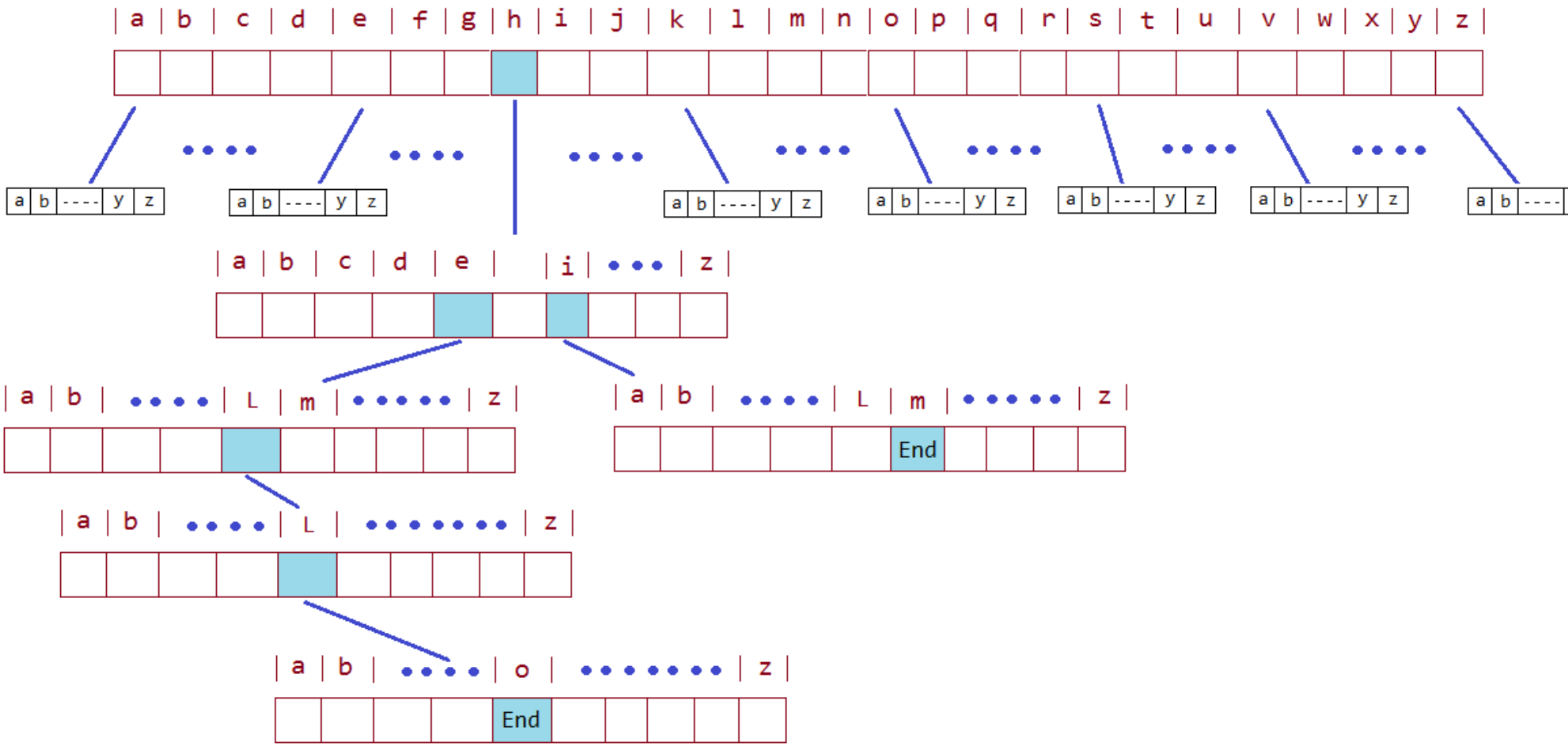
as posições vazias representam `NULL`.

Trie para "hello"



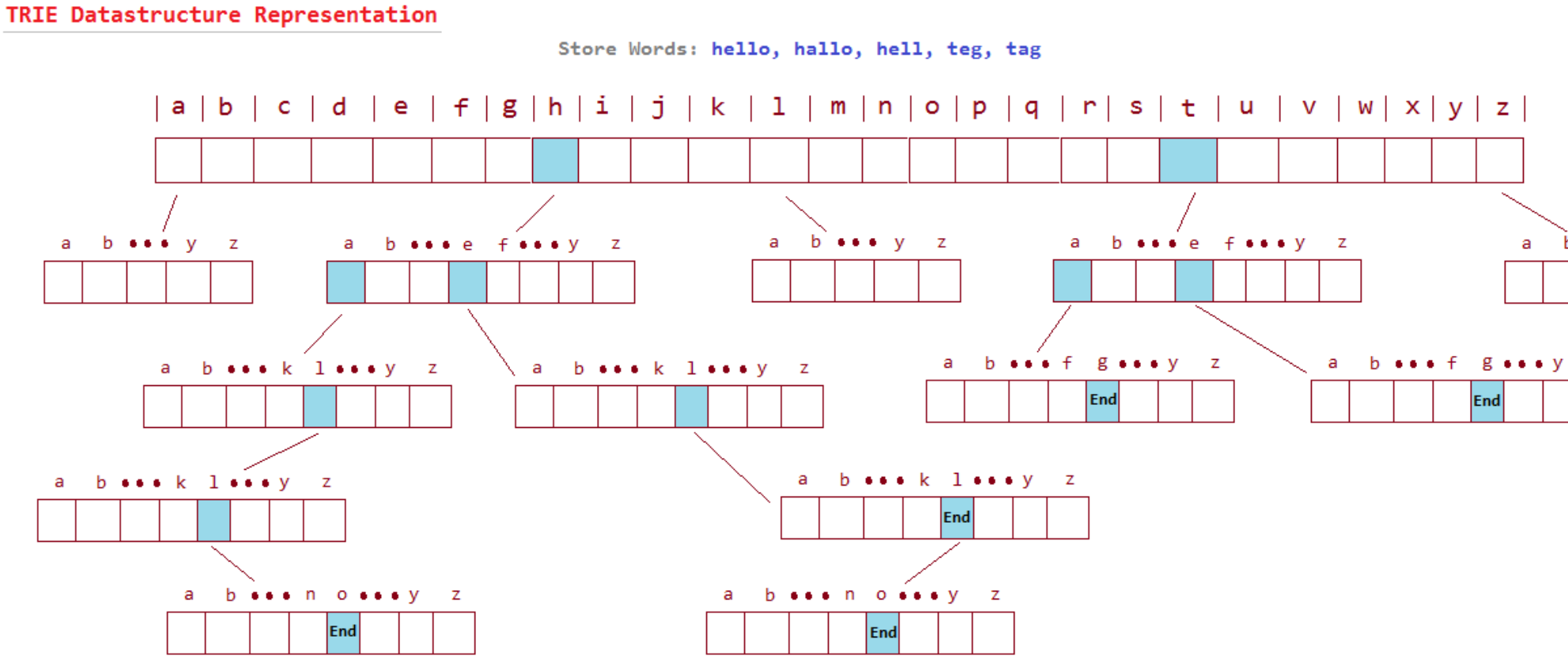
Fonte: JavaByPatel

Trie para "hello" e "him"



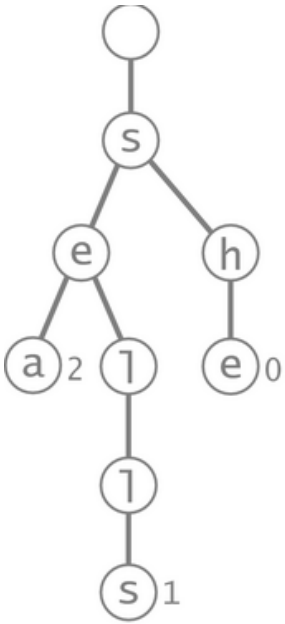
Fonte: JavaByPatel

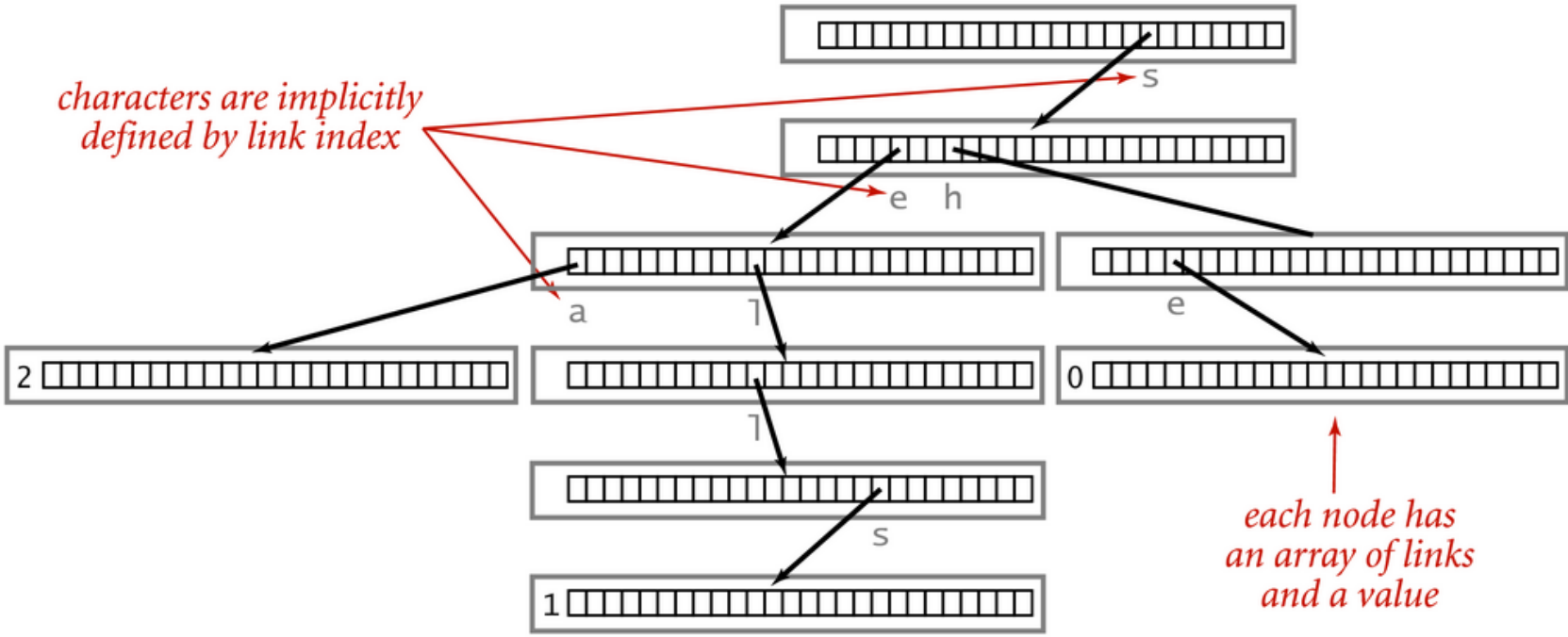
Trie para ...



Fonte: JavaByPatel

Trie para "sea", "sells" e "he"

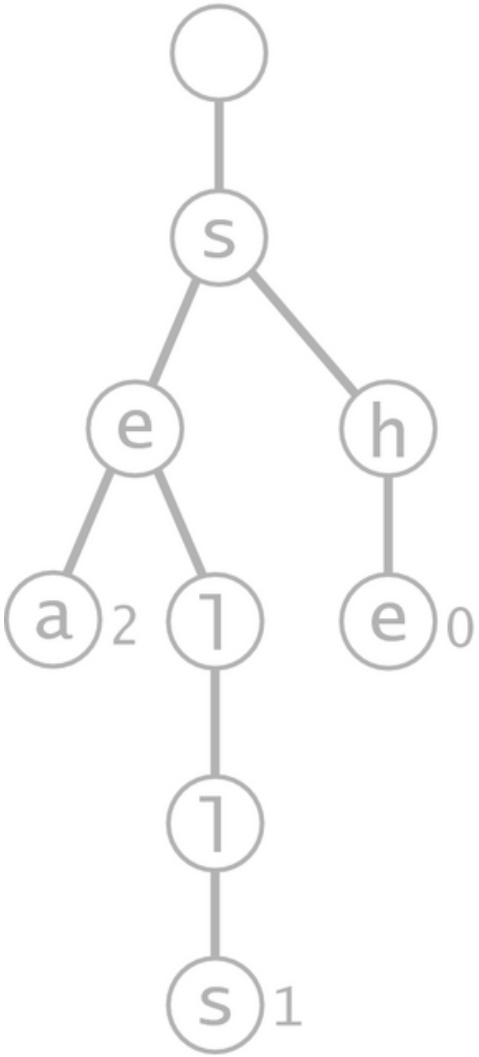


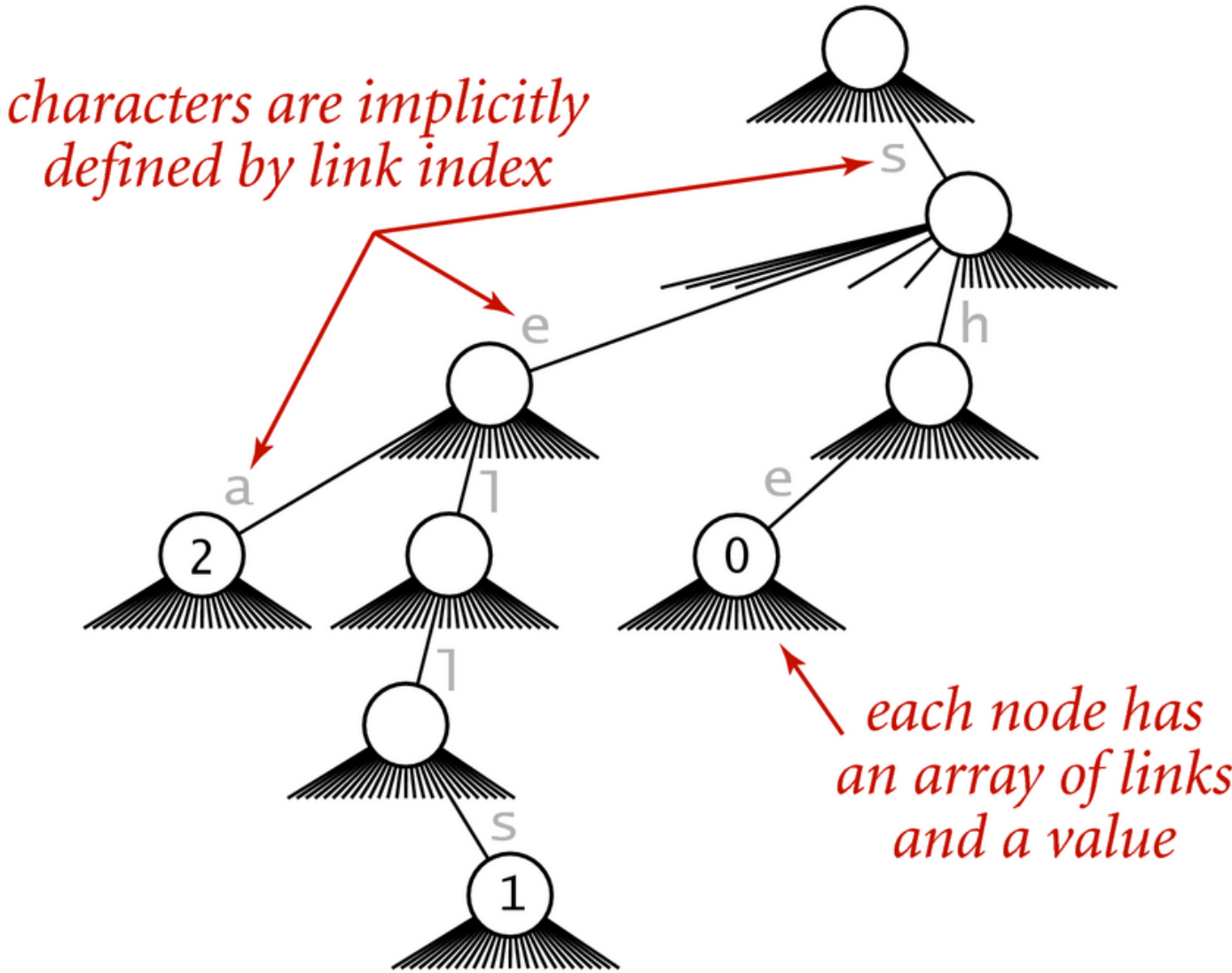


Trie representation ($R = 26$)

algs4

Outra representação da mesma trie

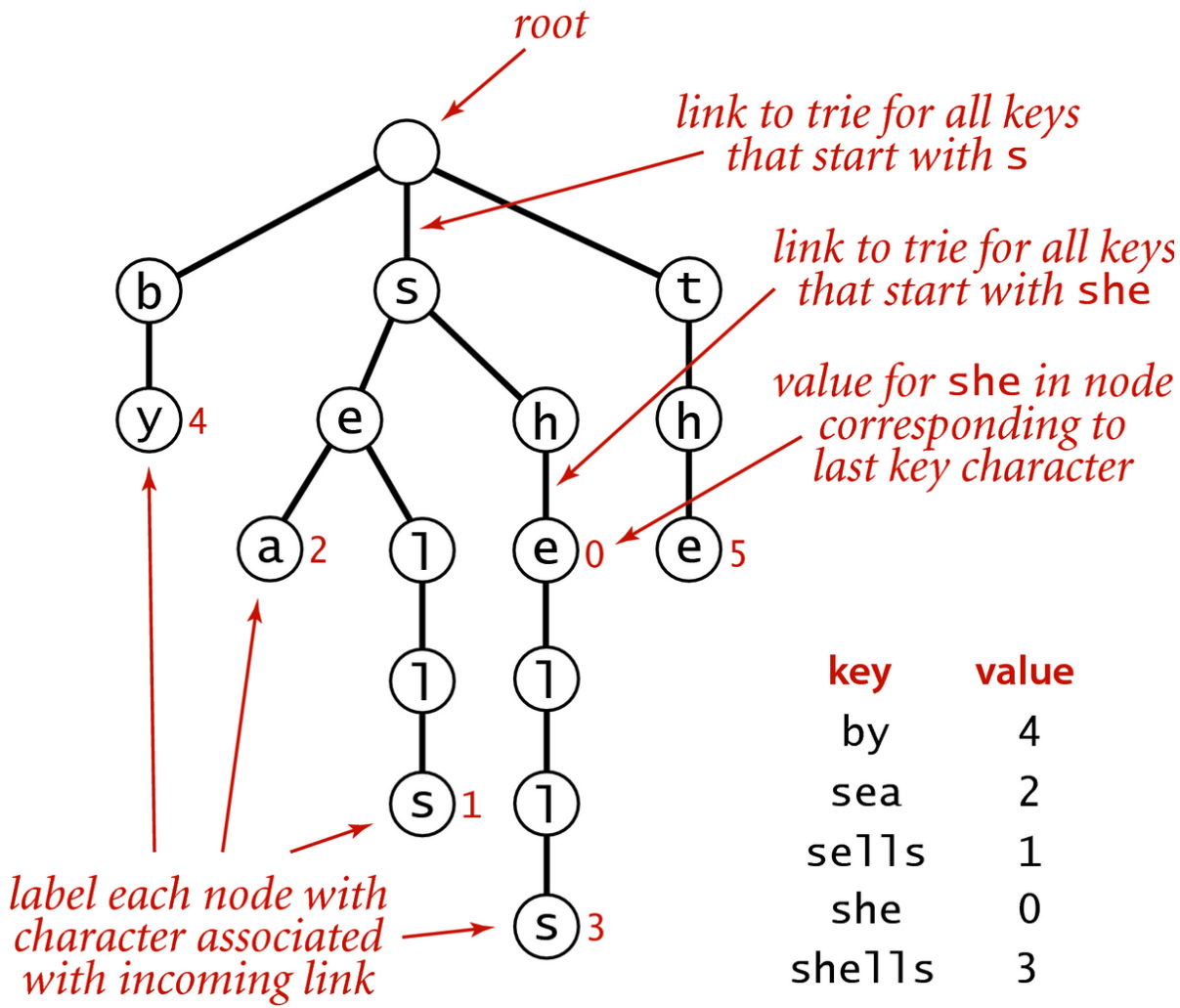




Trie representation

Fonte: `algs4`

Anatomia de uma Trie



key	value
by	4
sea	2
sells	1
she	0
shells	3
the	5

Anatomy of a trie

Fonte: algs4

Observações

Duas observações importantes sobre tries:

- chaves ficam codificadas nos caminhos que começam na raiz;
- prefixos de chaves, que nem sempre são chaves, estão representados na trie.

Ao descer da raiz até um nó `x`, **soletramos** uma string, digamos `s`. Dizemos que `s` leva ao nó `x`.

Dizemos também que o nó `x` é localizado pela string `s`.

A string que leva a um nó `x` é uma chave se e somente se `x->val != NULL`.

Alfabeto

Um **alfabeto** é um conjunto
de caracteres ou símbolos.

O tipo `enum` em C permite definir
um alfabeto personalizado com `R` caracteres,
numerados de 0 a `R`-1.

`R` é a base do alfabeto.

Cada unsigned char é um número entre 0 e 255.

Em nossas implementações, manipularemos strings,
então usaremos simplesmente o tipo char
(somente os números positivos, entre 0 e 127).

Alfabeto do DNA

Alfabeto: A, C, T, G;

```
enum nucleotideos {A, C, T, G} Nucleotideo;
```

c	A	C	T	G
valor	0	1	2	3

i	0	1	2	3
Nucleotideo	A	C	T	G

Alfabeto ASCII

ASCII Code Chart

	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
0	NUL	SOH	STX	ETX	EOT	ENQ	ACK	BEL	BS	HT	LF	VT	FF	CR	SO	SI
1	DLE	DC1	DC2	DC3	DC4	NAK	SYN	ETB	CAN	EM	SUB	ESC	FS	GS	RS	US
2		!	"	#	$	%	&	'	(	)	*	+	,	-	.	/
3	0	1	2	3	4	5	6	7	8	9	:	;	<	=	>	?
4	@	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
5	P	Q	R	S	T	U	V	W	X	Y	Z	[	\	]	^	_
6	`	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
7	p	q	r	s	t	u	v	w	x	y	z	{	|	}	~	DEL

Fonte: Wikipedia

TrieST.c: esqueleto

```
static int R = 128;                /* tamanho do alfabeto */
static int n;              /* número de pares chave-valor */

/* definição e rotinas de Node */
{...}
static Node r; /* raiz da trie */

Value  get(char *key) {...}
static Node getT(Node x, char *key, int d) {...}
void   put(char *key, Value val) {...}
static Node putT(Node x, char *key,
                 Value val, int d) {...}
void   delete(char *k) {...}
Queue  keys() {...}   /* Iterador */
```

get(key): método clássico

Seguimos os ponteiros soletrando a string key.

```
Value get(char *key) {
   Node x = getT(r, key, 0);
   if (x == NULL) return NULL;
   return x->val;
}
```

O terceiro parâmetro em getT é
o quanto já soletramos na busca.

get(key): método clássico

Seguimos os ponteiros soletrando a string key.

```
static Node getT(Node x, char *key, int d)
{
   if (x == NULL) return NULL;
   if (d == strlen(key)) return x;
   char c = key[d];
   return getT(x->next[c], key, d+1);
}
```

O terceiro parâmetro em getT é
o quanto já soletramos na busca.

put(key, val): método clássico

É feita uma busca.
Se a key é encontrada, o valor val é substituído.
Caso contrário chegamos a um NULL
e devemos continuar a inserção ou
chegamos no último caractere de key.

```
void put(char *key, Value val) {
   r = putT(r, key, val, 0);
}
```

put(key, val): método clássico

```
static Node putT(Node x, char *key,
                 Value val, int d) {
   if (x == NULL) x = newNode();
   if (d == strlen(key)) {
      if (x->val == NULL) n++;
      x->val = val;
      return x;
   }
   char c = key[d];
   x->next[c] = putT(x->next[c],key,val,d+1);
   return x;
}
```

newNode(): inicializa com NULL os apontadores.

`delete(key)`: método clássico

`delete(key)` remove a chave `key`
do conjunto de chaves da trie.
Em princípio, a implementação de `delete()` é fácil:
basta encontrar o nó `x` localizado por `key` e fazer

```
x->val = NULL;
```

Infelizmente, a trie resultante dessa operação
pode não ser *limpa*.

Para manter a `trie` limpa,
é preciso fazer algo mais complexo.

delete(key): método clássico

```
void delete(char *key) {
   r = deleteT(r, key, 0);
}
```

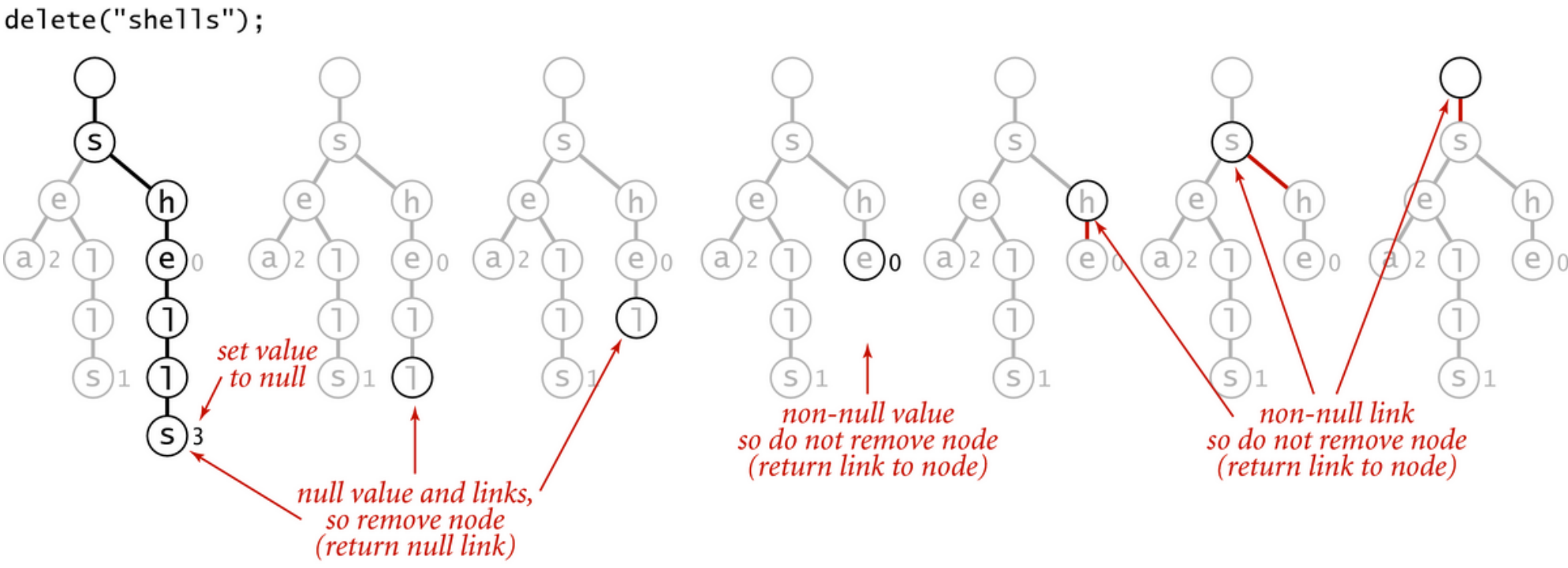
delete(key): método clássico

```
static Node deleteT(Node x, char *key, int d)
{
   if (x == NULL) return NULL;
   if (d == strlen(key)) x->val = NULL;
   else {
      char c = key[d];
      x->next[c] = deleteT(x->next[c],key,d+1);
   }
```

delete(key): método clássico

```
static Node deleteT(Node x, char *key, int d)
{
   if (x == NULL) return NULL;
   if (d == strlen(key)) x->val = NULL;
   else {
      char c = key[d];
      x->next[c] = deleteT(x->next[c],key,d+1);
   }
   if (x->val != NULL) return x;
   for (char c = 0; c < R; c++)
      if (x->next[c] != NULL) return x;
   return NULL;
}
```

delete(key): ilustração



Deleting a key (and its associated value) from a trie

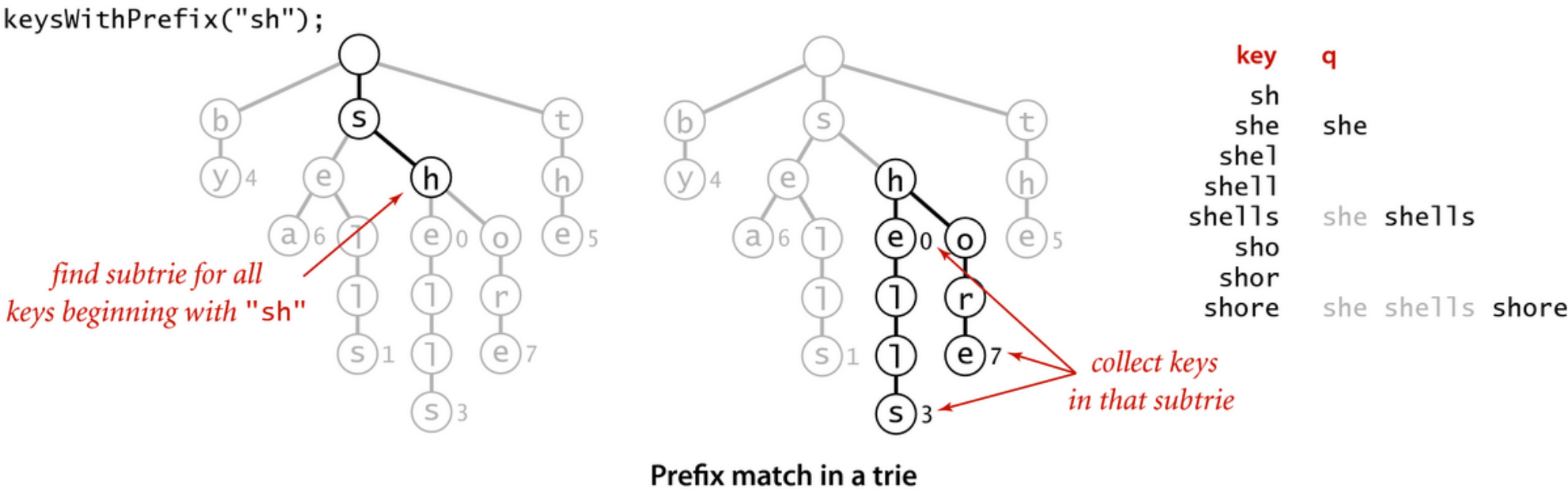
Fonte: algs4

keysWithPrefix(pre): método especial

Devolve todas as chaves na ST que têm prefixo pre.

```
Queue keysWithPrefix(char *pre) {
   Queue q = queueInit();   /* fila de chaves */
   Node x = getT(r, pre, 0);
   collect(x, pre, q);
   return q;
}
```

`keysWithPrefix(pre)`: ilustração



Prefix match in a trie

Fonte: `algs4`

collect(): método auxiliar

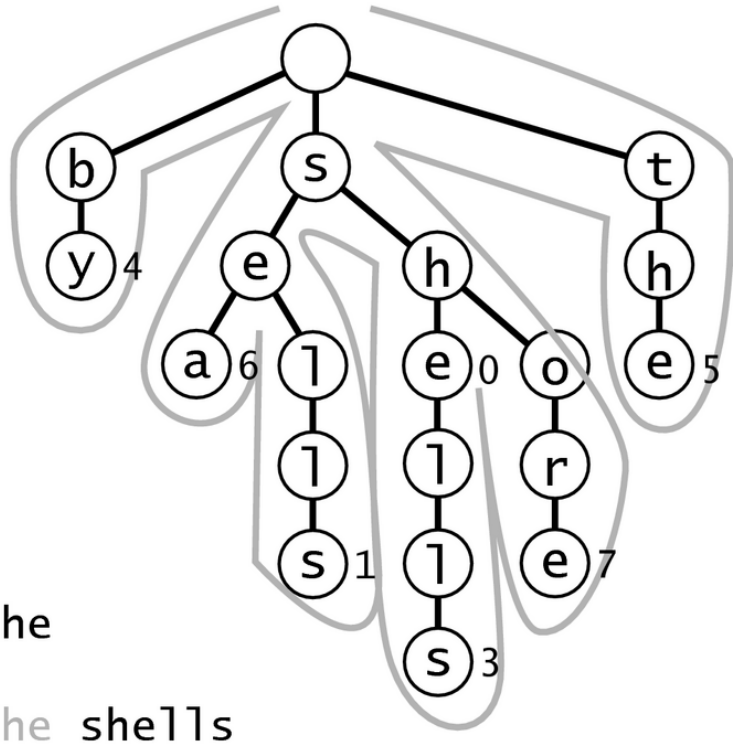
O método coloca na fila q todas as chaves da subtrie cuja raiz é x depois de acrescentar o prefixo pre a todas essas chaves.

```
static void collect(Node x, char *pre, Queue q) {
    int n = strlen(pre);
    if (x == NULL) return;
    if (x->val != NULL)
        enqueue(q, pre);      /* enfila uma cópia */
    pre[n+1] = '\0';
    for (char c = 0; c < R; c++)
        pre[n] = c;
        collect(x->next[c], pre, q);
    pre[n] = '\0';
}
```

collect(): ilustração

```
keysWithPrefix("");
```

key	q
b	
by	by
s	
se	
sea	by sea
sel	
sell	
sells	by sea sells
sh	
she	by sea sells she
shell	
shells	by sea sells she shells
sho	
shor	
shore	by sea sells she shells shore
t	
th	
the	by sea sells she shells shore the



Collecting the keys in a trie (trace)

Fonte: algs4

keys(): método clássico

```
Queue keys() {
   return keysWithPrefix("");
}
```

longestPrefixOf(s): método especial

Devolve a maior chave que é prefixo de s.

```
char *longestPrefixOf(char *s) {
    int max = -1; char *p;
    Node x = r;
    for (int d = 0; x != NULL; d++) {
        if (x->val != NULL) max = d;
        if (d == strlen(s)) break;
        x = x->next[s[d]];
    }
    if (max == -1) return NULL;
    p = mallocSafe((max+1)*sizeof(char));
    strncpy(p, s, max);
    p[max] = '\0';
    return p;
}
```

keysThatMatch(): método especial

Devolve todas as chaves que
casam com o padrão pat.

Os caracteres '.' em pat são curingas.

```
Queue keysThatMatch(char *pat) {
   Queue q = queueInit();
   collectC(r, "", pat, q);
   return q;
}
```

Mais um collect()

Coloca em q todas as chaves da trie que
têm prefixo pre e casam com o padrão pat.

```
static void
collectC(Node x, char *pre, char *pat, Queue q) {
    int n = strlen(pre);
    if (x == NULL) return;
    if (strlen(pre) == strlen(pat) && x->val != NULL)
        enqueue(q, pre);          /* enfila uma cópia */
    if (strlen(pre) == strlen(pat)) return;
```

Mais um `collect()`

Coloca em `q` todas as chaves da trie que têm prefixo `pre` e casam com o padrão `pat`.

```
    char c_next = pat[strlen(pre)];
    pre[n+1] = '\0';
    for (int c = 0; c < R; c++)
        if (c_next == '.' || c_next == c)
            pre[n] = c;
            collectC(x->next[c], pre, pat, q);
    pre[n] = '\0';
}
```

Consumo de espaço e tempo

A estrutura de uma `trie` não depende da ordem em que as chaves são inseridas ou removidas.

O consumo de tempo das operações sobre uma `trie` não depende do número `n` de itens.

O número de nós visitados por `get(key)` é no máximo $1 + w$, onde $w = $ `strlen(key)`.

O número de links em uma `trie` está entre `Rn` e `Rnw`, onde `w` é o comprimento médio de uma chave.

Consumo de espaço e tempo

O número esperado de nós visitados durante uma busca mal sucedida em uma `trie` com `n` chaves aleatórias sobre um alfabeto de tamanho `R` é aproximadamente $\log_{\texttt{R}} \texttt{n}$.

Tries ternárias

Fonte: `The Little Prince, Antoine de Saint-Exupéry`

Referências: Tries (árvores digitais) (PF); Tries (S&W); slides (S&W); Vídeo (S&W); TAOCP, vol 3, cap. 6.3

Tries ternárias

O maior problema das `tries` é possivelmente o espaço, já que cada nó contém `R` referências.

Assim, cada nó utiliza pelo menos $8 \times R$ bytes.

Veja a seguir alguns valores de `R` para alguns alfabetos comuns em aplicações.

Para evitar o custo excessivo de espaço de uma `R-trie`, consideramos uma representação como uma **ternary search trie** (`TST`).

Alphabet

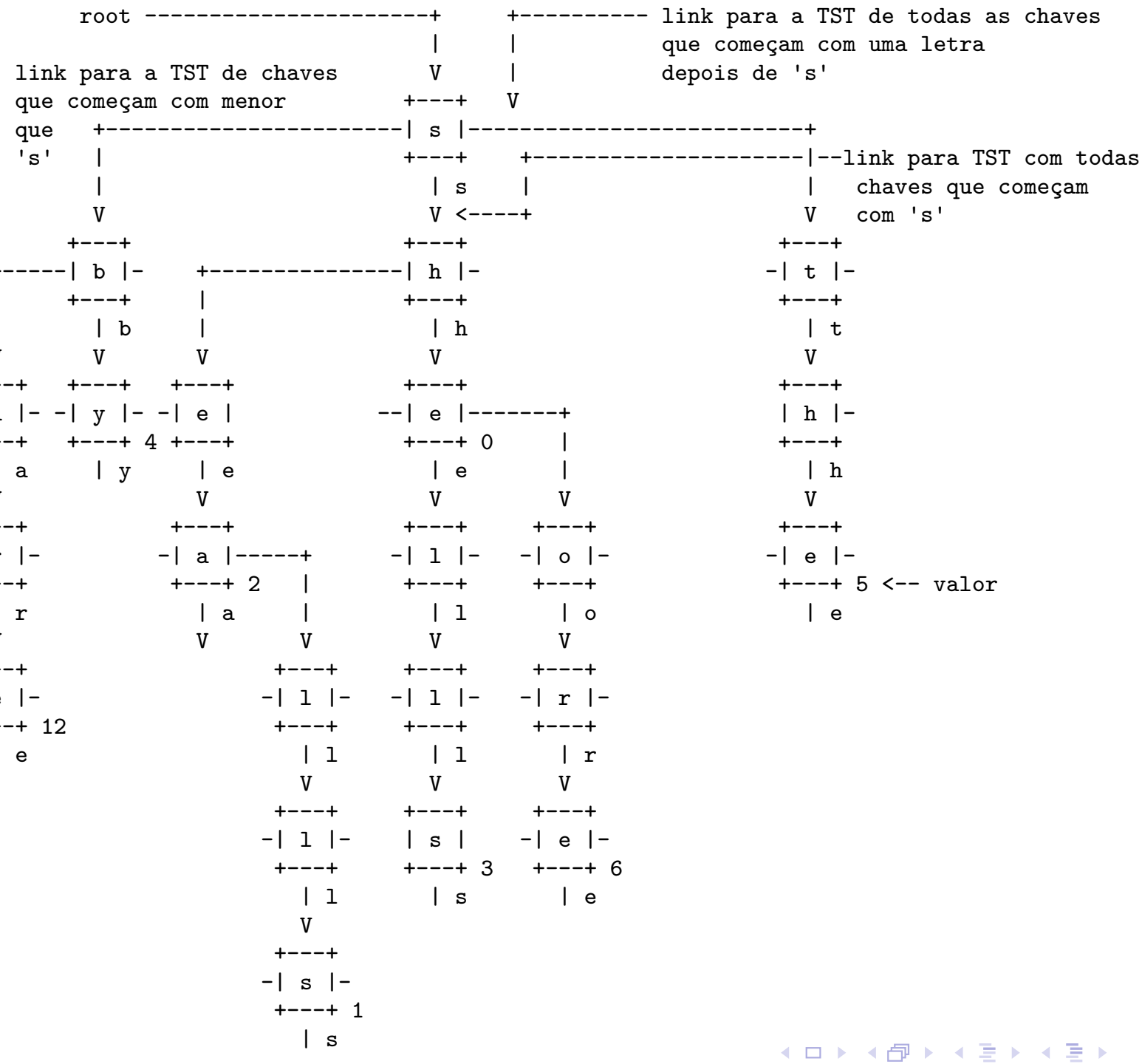
nome	R()	lg(R)	conjunto de caracteres
BINARY	2	1	'01'
DNA	4	2	'ACTG'
OCTAL	8	3	'01234567'
DECIMAL	10	4	'0123456789'
HEXADECIMAL	16	4	'0123456789ABCDEF'
PROTEIN	20	5	'ACDEFGHIKLMNPQRSTVWY'
LOWERCASE	26	5	'abcd...wxyz'
UPPERCASE	26	5	'ABCD...WXYZ'
ASCII	128	7	alfabeto ASCII
EXTENDED_ASCII	256	8	alfabeto ASCII estendido
UNICODE16	65536	16	alfabeto Unicode

Outra ilustração

Para os pares `key`-`val`

`key`	`val`
are	12
by	4
sea	2
sells	1
she	0
shells	3
the	5
shore	6

temos a `TST` a seguir.


root
link para a TST de chaves que começam com menor que 's'
link para a TST de todas as chaves que começam com uma letra depois de 's'
link para TST com todas chaves que começam com 's'
b
a
y
4
e
a
2
l
l
s
1
s
h
e
0
l
l
s
3
s
t
h
e
5 <-- valor
o
r
e
6
e
r
e
12
e

TSTs

De maneira semelhante ao que ocorre com `tries`, nas `tries` ternárias:

- chaves ficam codificadas nos caminhos que começam na raiz;
- prefixos de chaves, que nem sempre são chaves, estão representados na `TST`.

Estrutura de uma trie ternária

Os links da estrutura correspondem a caracteres.
Nas figuras, o caractere escrito dentro de um nó é o caractere do link que sai pelo meio do nó.
TSTs são compostas por nós do tipo Node.

```
typedef struct node *Node;

struct node {
   char c;      /* caracter */
   Value val;
   Node left;   /*  < c */
   Node mid;    /* == c */
   Node right;  /*  > c */
};
```

Ilustração

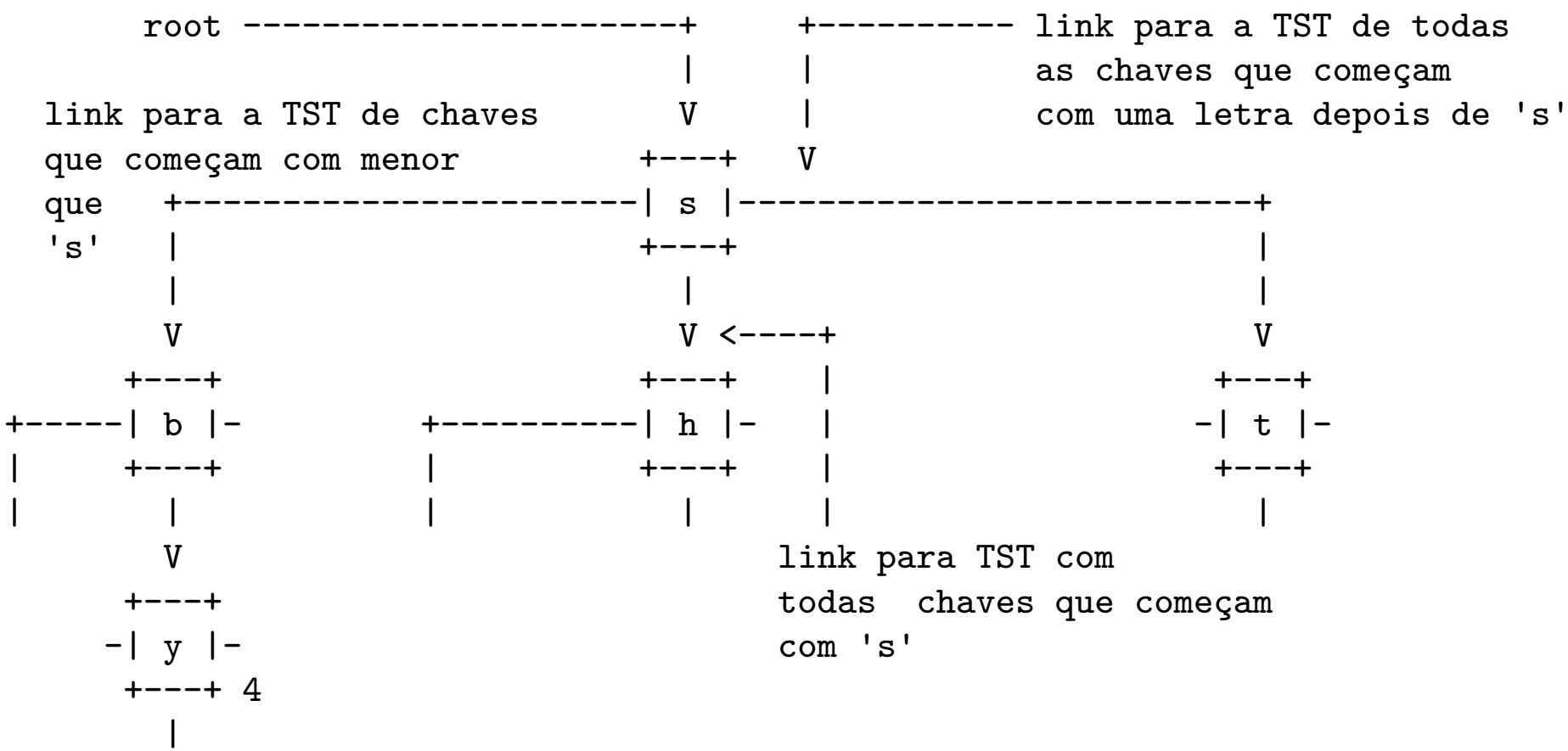
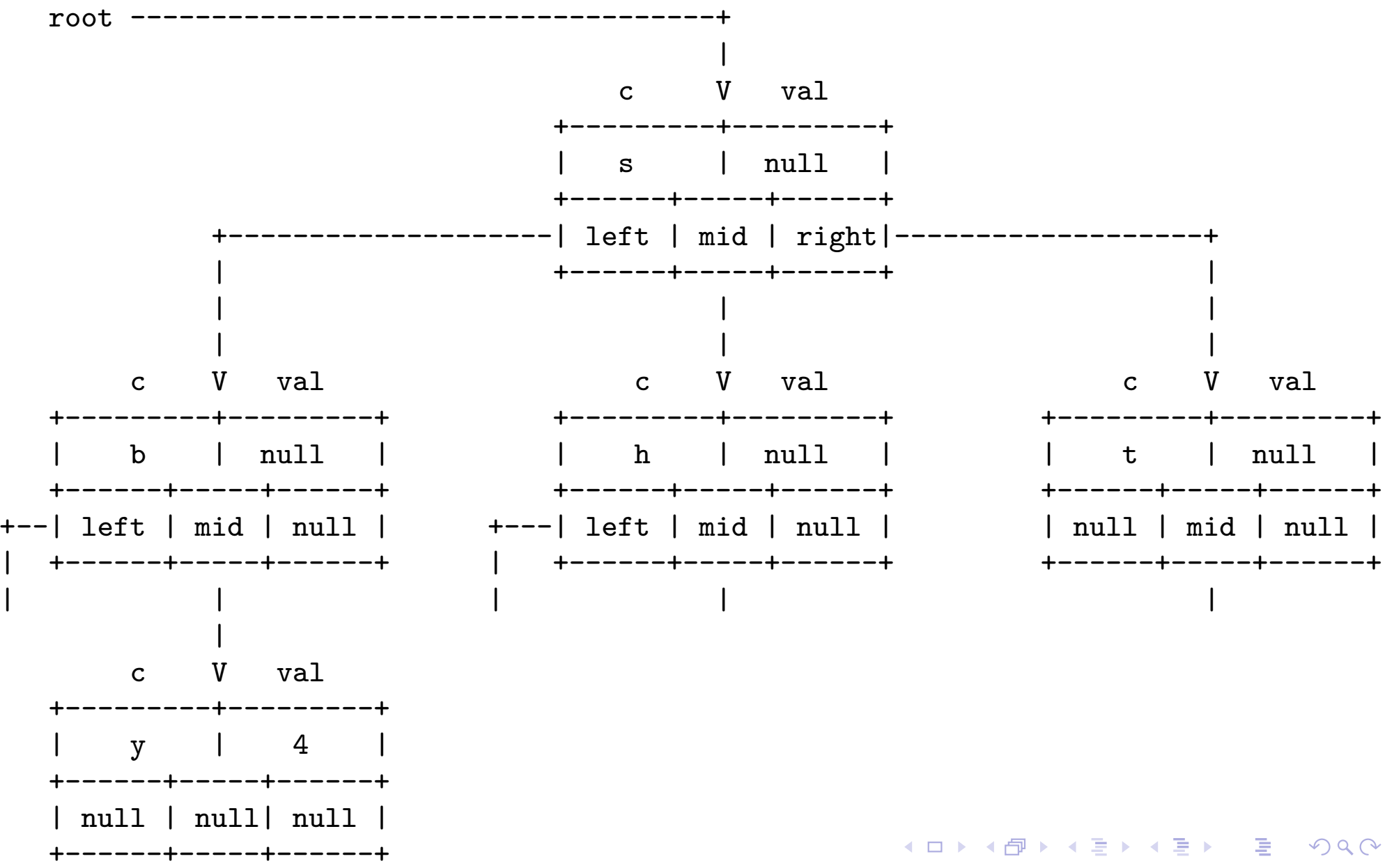
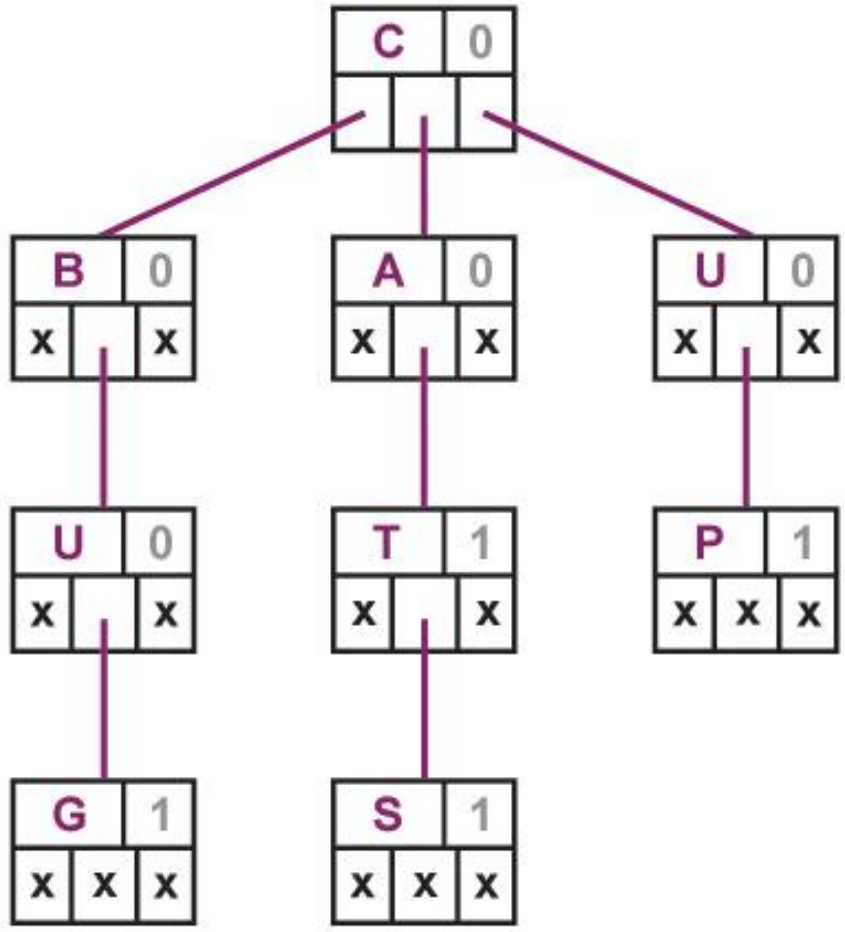

root
link para a TST de chaves que começam com menor que 's'
link para a TST de todas as chaves que começam com uma letra depois de 's'
s
b
h
t
y
4
link para TST com todas chaves que começam com 's'

Ilustração com zoom



Outra ilustração



Ternary Search Tree for CAT, BUG, CATS, UP

Fonte: Ternary search trees for autocompletion and spell checking

TST.c: esqueleto

```
static int R = 128;                /* tamanho do alfabeto */
static int n;                  /* número de pares chave-valor */

/* definição e rotinas de Node */
{...}
static Node r; /* raiz da tst */

Value  get(char *key) {...}
static Node getT(Node x, char *key, int d) {...}
void   put(char *key, Value val) {...}
static Node putT(Node x, char *key,
                 Value val, int d) {...}
void   delete(char *k) {...}
Queue  keys() {...}   /* Iterador */
```

get(key): método clássico

A string que leva a um nó x é uma chave se e somente se x->c é o último caractere da chave e x->val != NULL.

```
/* Trie e TST */
Value get(char *key) {
   Node x = getT(r, key, 0);
   if (x == NULL) return NULL;
   return x->val;
}
```

get(key): método clássico

```
static Node getT(Node x, char *key, int d) {
   char c = key[d];
   if (x == NULL) return NULL;
   if (c < x->c)
      return getT(x->left, key, d);
   if (c > x->c)
      return getT(x->right, key, d);
   if (d < strlen(key)-1)
      return getT(x->mid, key, d+1);
   return x;
}
```

put(key, val): método clássico

É feita uma busca.
Se a key é encontrada o valor val é substituido.
Caso contrário, chegamos a um NULL e devemos continuar a inserção, ou chegamos no último caractere de key.

```
/* Trie e TST */
void put(char *key, Value val) {
   r = putT(r, key, val, 0);
}
```

put(key, val): método clássico

```
static Node putT(Node x, char *key,
                 Value val, int d) {
    char c = key[d];
    if (x == NULL)
        { x = newNode(); x->c = c; }
    if (c < x->c)
        x->left = putT(x->left, key, val, d);
    else if (c > x->c)
        x->right = putT(x->right, key, val, d);
    else if (d < strlen(key)-1)
        x->mid = putT(x->mid, key, val, d+1);
    else x->val = val;
    return x;
}
```

collect(): método auxiliar

O método coloca na fila q todas as chaves da subtrie cuja raiz é x, depois de acrescentar o prefixo pre a todas essas chaves.

```
static void collect(Node x, char *pre, Queue q) {
    int n = strlen(pre);
    if (x == NULL) return;
    collect(x->left, pre, q);
    /* ordem lexicográfica */
    pre[n] = x->c;   pre[n+1] = '\0';
    if (x->val != NULL) enqueue(q, pre);
    collect(x->mid, pre, q);
    pre[n] = '\0';
    collect(x->right, pre, q);
}
```

keys(): método clássico

```
Queue keys() {
    Queue q = queueInit();
    collect(r, "", q);
    return q;
}
```

keysWithPrefix(): método especial

Devolve todas as chaves na ST que têm prefixo pre.

```
Queue keysWithPrefix(char *pre) {
   Queue q = queueInit();   /* fila de chaves */
   Node x = getT(r, pre, 0);
   if (x == NULL) return q;
   if (x->val != NULL) enqueue(q, pre);
   collect(x->mid, pre, q);
   return q;
}
```

`longestPrefixOf()`: método especial

Devolve a maior chave que é prefixo de `s`.

```
char *longestPrefixOf(char *s) {
   int max, i; char c, *p;
   Node x = r;

   if (s == NULL || strlen(s) == 0)
      return NULL;

   max = i = 0; c = s[0];
```

longestPrefixOf(): método especial

```
    while (x != NULL && i < strlen(s)) {
       if (c < x->c) x = x->left;
       else if (c > x->c) x = x->right;
       else {
          i++;
          if (x->val != NULL) max = i;
          x = x->mid;
       }
    }
    p = mallocSafe((max+1)*sizeof(char));
    strncpy(p, s, max);  p[max] = '\0';
    return p;
}
```

keysThatMatch(): método especial

Devolve todas as chaves que
casam com o padrão pat.

Os caracteres '.' em pat são curingas.

```
Queue keysThatMatch(char *pat) {
   Queue q = queueInit();
   collectC(r, "", 0, pat, q);
   return q;
}
```

collectC: versão do collect com curinga.

Mais um collect()

Coloca em q todas as chaves da trie que
têm prefixo pre e casam com o padrão pat.

```
static void collectC(Node x, char *pre,
                     int i, char *pat, Queue q) {
   int n = strlen(pre);
   char c = pat[i];

   if (x == NULL) return;
   if (c == '.' || c < x->c)
      collectC(x->left, pre, i, pat, q);
```

Mais um collect()

```
    pre[n] = x->c;   pre[n+1] = '\0';
    if (c == '.' || c == x->c) {
       if (i == strlen(pat) - 1
          && x->val != NULL)
          enqueue(q, pre);
       else if (i < strlen(pat) - 1)
          collectC(x->mid, pre, i+1, pat, q);
    }
    pre[n] = '\0';
    if (c == '.' || c > x->c)
       collectC(x->right, pre, i, pat, q);
}
```

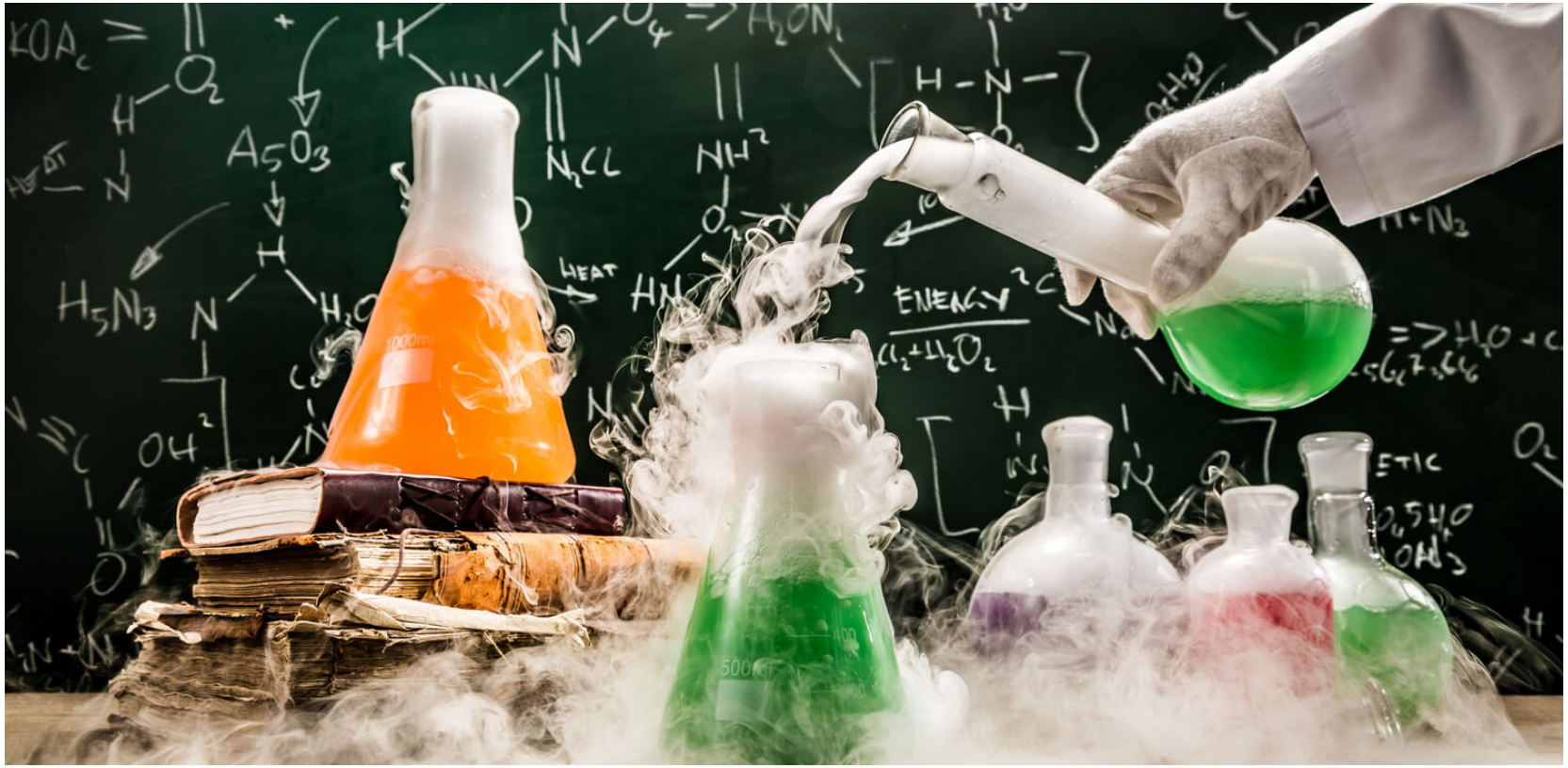
Consumo de espaço e tempo

Espaço. A propriedade mais importante de uma TST é que ela tem apenas três links por nó.

Proposição J: O número de links em uma TST com n chaves de comprimento médio w é entre 3n e 3nw.

Proposição K: O número esperado de nós visitados durante uma busca mal sucedida em uma TST com n chaves aleatórias é aproximadamente $\lg n$.

Alguns experimentos



Fonte: https://singularityhub.com/

Experimentos: les_miserables.txt

ST com 26764 itens

```
% wc les-miserables.txt
68116 568531 3322649 les-miserables.txt
```

ST	criada (s)
BST	0.626
RedBlackBST	0.577
SeparateChainST	0.495 (4096, 18, $\alpha = 6$)
LinearProbingST	0.411 (65536, 67, $\alpha = 0.4$)
TrieST	0.574
TST	0.497

Experimentos: actors.list

ST com 1482495 itens

```
% wc actors.list
16612200 124796815 932688622 actors.list
```

ST	criada (s)
BST	141.968
RedBlackBST	168.723
SeparateChainST	110.035 (262144, 19, $\alpha = 5$)
LinearProbingST	73.123 (4194304, 4787, $\alpha = 0.35$)
TrieST	OutOfMemoryError
TST	107.223