Elementi di Bioinformatica

Gianluca Della Vedova

Univ. Milano–Bicocca
http://gianluca.dellavedova.org

10 ottobre 2019

Trie

Trie

- Albero
- Query: parola ∈ dizionario
- archi etichettati
- Percorso radice-foglia = parola

Dizionario

ABRACADABRA
ARRAY
ABRA

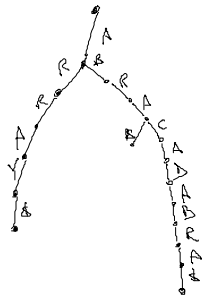
Trie

Terminatore

$ non appartiene all'alfabeto

Dizionario

ABRACADABRA$
ARRAY$
ABRA$



Suffix tree

Definizione

- Trie compatto di tutti i suffissi di T\$
- Le etichette degli archi uscenti da x iniziano con simboli diversi
- suffissi ⇔ percorso radice-foglia

BANANA$

Suffix tree 2: Definizione

- foglie etichettata con posizione inizio suffisso
- path-label(x): concatenazione etichette
- string-depth(x): lunghezza path-label(x)
- Pattern matching = visita

Problemi

- Spazio $O(n^2)$
- Puntatori al testo (posizioni)
- Spazio $20n$ bytes

Suffix array

Definizione

- Array dei suffissi in ordine lessicografico
- Posizioni iniziali del suffisso nell'array
- Spazio $4n$ bytes
- $Lcp[i]$: lunghezza prefisso comune $SA[i]$, $SA[i+1]$

BANANA$

i	1	2	3	4	5	6	7
SA	7	6	4	2	1	5	3
Lcp	0	1	3	0	0	2	-

Da Suffix tree a Suffix array

- Visita depth-first di ST
- archi uscenti di ogni nodo in ordine lessicografico
- $Lcp[i]$ = string-depth di $lca(i, i+1)$

Da Suffix array a Suffix tree

- $Lcp = 0$: partizione SA
- corrispondono ai figli della radice
- ricorsione prendendo i

Suffix tree generalizzato

s_1:
BANANA$
s_2:
PANNA$

Sottostringa comune più lunga di due stringhe

Due stringhe s_1 e s_2

- Suffix tree generalizzato = insieme di stringhe
- $ST(s_1\$_1 s_2\$_2)$
- Nodo x con foglie di s_1 e s_2
- Sottostringa di s_1 e s_2
- $ST(s_1\$ s_2\$)$
- Max string-depth

Licenza d'uso