

Advanced string algorithms

Jill-Jênn Vie

October 17, 2023

Reminder: Knuth-Morris-Pratt

Let s a string of length n

Prefix function

Array p such that $p[i]$ is the length of the longest proper prefix of $s[0..i]$ which is also a suffix of $s[0..i]$.

Idea

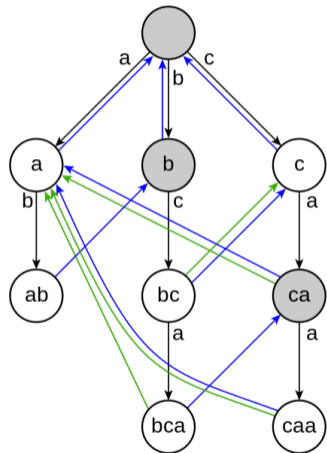
Build p in $O(n)$ by dyn prog

Generalization

But now, how to find **all** occurrences of a set of patterns in a string?

Aho-Corasick

Look for all occurrences of a, ab, bc, bca, c, caa (white nodes)



Blue arrows: suffix links

Green arrows: terminal

Complexity

If \sum strings is m , nb vertices n , alphabet size k

- ▶ $O(mk)$ thanks to dyn prog
- ▶ Can be sped up $O(n \log k)$ with a segment tree

Note

- ▶ Generalization of KMP for several strings
- ▶ Notebook implementation is exactly cp-algorithms (wtf 445 pages of Stanford slides)

Problems using Aho-Corasick

- ▶ Find all strings from a given set in a text
- ▶ Finding the lexicographical smallest string of a given length that doesn't match any given strings
- ▶ Finding the shortest string containing all given strings
- ▶ Finding the lexicographical smallest string of length L containing k strings

Rabin-Karp: hashing

Looking for s in t

Idea

Comparing an updated rolling hash of every substring of t of size $|s|$ with the hash of s .

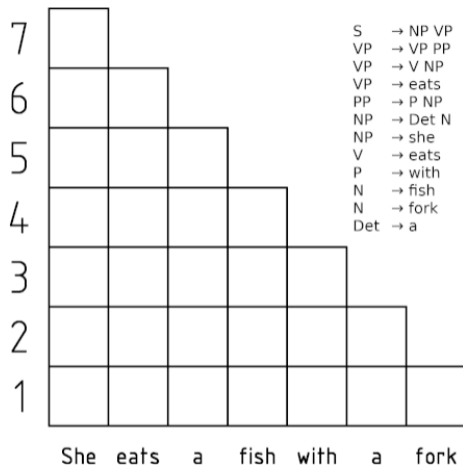
$$\text{hash}(x) = \sum_i x[i]A^i$$

Other application

- ▶ Lexicographical smallest rotation of a string: how to do it?

Recognize a context-free grammar

$S \rightarrow NP VP$
 $VP \rightarrow VP PP$
 $VP \rightarrow V NP$
 $VP \rightarrow \text{eats}$
 $PP \rightarrow P NP$
 $NP \rightarrow \text{Det } N$
 $NP \rightarrow \text{she}$
 $V \rightarrow \text{eats}$
 $P \rightarrow \text{with}$
 $N \rightarrow \text{fish}$
 $N \rightarrow \text{fork}$
 $\text{Det} \rightarrow \text{a}$



Complexity of CYK algorithm

$O(n^3 |G|)$ for string of length n and grammar of size $|G|$

Homework (several valid solutions): SWERC 2014's J: The Big Painting

```
XXXXXXOXXO
OXXO OOXOOX
XOOX XX XOOX
XOOX XX OXXO
OXXO XXXXXX
OOOOXXXXXX
XXXOXXOXXO
OOOXOOXOOX
OOOXOOXOOX
XXXOXXOXXO
```

<https://open.kattis.com/problems/bigpainting>

String data structures

	String Hashing	Suffix Array	Aho-Corasick
Search for duplicate strings in array of strings	X		
Fast hash calculation of substrings of string	X		
Number of substrings of given string	X	X	
Finding smallest cyclic shift		X	
Finding substring in a string		X	
Comparing substrings of a string $a < b$		X	
Longest common prefix of substrings		X	
Find strings of given set in a text			X
Finding lexicographically smallest string that doesn't match			X
Finding smallest string that contains all given strings			X
Finding lexicographically smallest string that contains k strings			X