3/17/10

15-750

String Matching Using FFT

Let S, T be 2 strings of same length.

Def $H(S,T)$ = # of mismatches (Hamming Dist)

$\mathrm{Conv}_H(P,S) \equiv H$ a string of length $n-m+1$

$$H_i = H(P, S_i^{m+i-1})$$

Open: Find an $O(n \log m)$ alg for $\mathrm{Conv}_H(P,S)$

Assume $\Sigma \subseteq \mathbb{Z}$

$$(S,T) = \sum_{i=1}^{n} S_i T_i \equiv S \oplus T$$

Def: $\mathrm{Conv}_{(,)}(P,S)$

Using FFT $\mathrm{Conv}_{(,)}$ is $O(n \log n)$ time

Can you see $O(n \log m)$ time?

Consider $\langle S, T \rangle = \sum_{i=1}^{m} (S_i - T_i)^2$

<u>note!</u> $\langle S, T \rangle = 0$ iff $S = T$

$\therefore$ $\text{Conv}_{\langle , \rangle}$ solves string matching

---

$$\text{Conv}_{\langle , \rangle}(P, S)_i = \sum_{j=1}^{m} \left(P_j - S_{i+j-1}\right)^2 = \sum_{j=1}^{m} P_j^2 - 2 \sum_{j=1}^{m} P_j \cdot S_{i+j-1} + \sum_{j=1}^{m} S_{i+j-1}^2$$

$\sum P_j^2$ = constant

$\sum P_j S_{i+j}$ = FFT conv!

$\tau_i = \sum_{j=1}^{m} S_{i+j}^2$ = $O(n)$ time using sliding window. over all i.

$$S_i^2 + \cdots + S_{i+m-1}^2$$
$$\underbrace{S_{i+1}^2 + \cdots + S_{i+m}^2}$$

ie $\tau_{i+1} = \tau_i - S_i^2 + S_{i+m}^2$

String Matching with Wildcards

Input: no wild cards

Σ alphabet $|\Sigma|$ large

P pattern $|P| = m$

T text $|T| = n$

$*$ $\equiv$ wildcard ($*$ matches everything)

$\Sigma = \{1, \ldots, k\} \cup \{*\}$

<u>Recall</u> $\Sigma \subseteq \mathbb{Z}$

$S \odot T = \sum_{i=1}^{n} S_i \cdot T_i$

Then $\text{Conv}_{\odot}(P, T) \equiv C \qquad C_i = P \odot T_i^{m+i-1}$

is $O(n \log m)$ Time using FFT.

Two needed convolutions (wild cards)

$\langle S,T \rangle_{\text{Char}} \equiv$ # of char·char matches.

eg $S = a * b * c$ $\quad \langle S,T \rangle = 2$

$T = c\, c * * d$

1 $\qquad$ 1

Goal: Compute $\text{Conv}_{\text{Char}}(P,T)$

$$S_i' = \begin{cases} 1 & \text{if } S_i \neq * \\ 0 & \text{o.w} \end{cases} \qquad T_i' = \begin{cases} 1 & \text{if } T_i \neq * \\ 0 & \text{o.w} \end{cases}$$

Claim $\langle S,T \rangle_{\text{Char}} = S' \oplus T'$

$$\text{Conv}_{\text{Char}}(P,T) = \text{Conv}_{\oplus}(P',T')$$

Yet another Product Rule

$x, y \in \mathbb{Z}$

$$(x,y)_{CH} \equiv \begin{cases} 0 & \text{if } x=0 \text{ or } y=0 \\ \frac{x}{y}+\frac{y}{x}-2 & \text{o.w.} \quad \left(\frac{(x-y)^2}{xy}\right) \end{cases}$$

$$(S,T)_{CH} \equiv \sum_{i=1}^{n} (s_i,t_i)_{CH}$$

Def $\mathrm{Corr}_{CH}(P,T)$

---

note $\frac{x}{y}+\frac{y}{x}-2 = \frac{x^2+y^2-2xy}{xy} = \frac{(x-y)^2}{x \cdot y}$

lemma:

$$(x,y)_{CH} = \begin{cases} 0 & \text{if } x=0 \text{ or } y=0 \text{ or } x=y \\ \geq \frac{1}{k(k-1)} & \text{o.w.} \end{cases} \qquad \Sigma = \{0,1,\ldots,k\}$$

$\Sigma = \{1, \ldots, k\}$ set $*$ to zero

Lemma $\mathrm{Conv}_{CH}(P, T)_i = 0$ iff

P matches T_i^{i+m-1} with wildcards

Computing Conv_{CH}

Suppose $P = a\, b * c, \ldots$

$P'' \equiv a, 1/a, b, 1/b, 0, 0, c, 1/c, \ldots$

Suppose $T = b, *, c, b, \ldots$

$T'' = (1/b, b, 0, 0, 1/c, c, 1/b, b, \ldots$

Lemma $\mathrm{Conv}_{\oplus}(P'', T'')$ (keeping only odd index values)

$= \mathrm{Conv}_{op}(P, T)$ where

$$(x,y)_{op} = \begin{cases} 0 & \text{if } x=0 \text{ or } y=0 \\ \frac{x}{y}+\frac{y}{x} & \text{o.w.} \end{cases}$$

Recall: $$(x,y)_{Char} = \begin{cases} 1 & \text{if } x \,\&\, y \text{ are charactors} \\ 0 & \text{o.w.} \end{cases}$$

Note $(x,y)_{CH} = (x,y)_{op} - 2(x,y)_{Char}$

$$(S,T)_{CH} = (S,T)_{op} - 2(S,T)_{Char}$$

$$\mathrm{Conv}_{CH}(P,T) = \mathrm{Conv}_{op}(P,T) - 2\,\mathrm{Conv}_{Char}(P,T)$$

$\therefore$ Conv_{CH} is $O(n \log m)$ Time