Algorithms
Professor John Reif

ALG 4.1

Randomized Pattern Matching:

Reading Selection:

CLR: Chapter 12

Handout: R. Karp and M. Rabin, "Efficient Randomized Pattern-Matching"

Generalized Pattern Matching

input index set R
for each r ε R strings X(r), Y(r) $\varepsilon \{0,1\}^m$

problem find r ε R s.t. X(r) = Y(r)

Examples:

(1) *ID String Pattern Matching*

input pattern $X = X_1 \ldots X_m \in \{0,1\}^m$

text $Y = Y_1 \ldots Y_n \in \{0,1\}^n$

index set R = {1,2,..., n-m+1}

$\forall r \in R$ $X(r) = X$

$Y(r) = Y_r\, Y_{r+1} \ldots Y_{r+(m-1)}$

(2) 2D Array Matching

input pattern

$s \times s$ binary array $X = (X_{ij})$, $m = s^2$

text $b \times b$ binary array $Y = (Y_{ij})$, $n = b^2$

index set $R = \{< i,j > \mid s \le i,j \le b\}$

$X(< i,j >)$ = string of rows of X

$Y(< i.j >)$ = string of rows in $s \times s$ block of Y with (i,j) in lower right postion.

(*note* Karp & Pratt reverse n,m)

Pattern Matching by Fingerprinting

S is a finite set

$\forall p \varepsilon S$ $\Phi_p(\cdot)$ is function $\{0,1\}^m \rightarrow$ small range D_p

$\Phi_p(X)$ is "*fingerprint*" for string X

idea compare $X(r) = Y(r)$ only if fingerprints agree: $\Phi_p(X(r)) = \Phi_p(Y(r))$

Algorithm

$p \leftarrow$ random element of S

for each r ε R in order *do*

begin

compute $a_p(r) = \Phi_p(X(r))$

compute $b_p(r) = \Phi_p(Y(r))$

if $a_p(r) = b_p(r)$ then

if $X(r) = Y(r)$ *then* output "Match at r"

end

fingerprint fn Φ_p : $\{0,1\}^m \rightarrow D_p$

Requirements

(1) small domain D_p

(2) small probability of *false match*

$\Phi_p(X(r)) = \Phi_p(Y(r))$ but $X(r) \neq Y(r)$

(3) fingerprints $\Phi_p(X(r))$, $\Phi_p(Y(r))$ are *easily updatable* from previous r

Examples of fingerprints:

(A)integer modular fns.

(B) unimodular matrices

(C) irreducible polynomial modular fns.

represent *binary string* $X = X_1 \ldots X_m$

by integer $H(x) = \sum_{i=1}^{m} X_i \; 2^{m-i}$

modular fingerprint $\Phi_p(x) = \mathrm{res}\,(H(x), p)$

modular fingerprint $\Phi_p(X) = \mathrm{res}\,(H(x), p)$

$$= p \cdot \left\lfloor \frac{H(X)}{p} \right\rfloor - H(X)$$

note $\Phi_p(x) \equiv H(X) \bmod p$

Define $S = \{p \mid p \text{ is prime and } p \le M\}$

where M is a (suf. large) integer

idea choose *random* $p \in S$

$\Rightarrow$ must prove $\Phi_p(X) = \text{res}(H(X), p)$ is good fingerprint

Facts about Prime Numbers

let $\Pi(k)$ = number of primes $\le k$

FACT 1 If $k \ge 29$ and $a \le 2^k$, then a has $\le \Pi(k)$ distinct prime divisors

proof follows from Rosser & Schoenfeld bound:

$$\prod_{\substack{p \text{ prime} \\ p \le k}} p > e^{k - 2.05282 k^{\frac{1}{2}}}$$

FACT 2 for all $k \ge 17$

$$\frac{k}{\ln k} \le \Pi(k) \le 1.25506 \frac{k}{\ln k}$$

(prime number theorem)

Suppose

Randomized Pattern Match

Algorithm is executed, with

fingerprint $\Phi_p(X) = res(H(x), p)$

with set $S = \{p \mid p \text{ prime} \le M\}$

where $M = mn^2$ and $mn \ge 29$

Theorem

for each $r \in R$, probability of

false match $\Phi_p(X(r)) = \Phi_p(Y(r))$ but $X(r) \ne Y(r)$

is $\le \dfrac{2.511}{n}$

proof

A false match occurs only if $\exists\, r \in R$

$X(r) \ne Y(r)$ and $P \mid (H(x(r)) - H(Y(r)))$

iff $p \mid L$ where $L = \prod_{X(r) \ne Y(r)} | H(X(r)) - H(Y(r)) |$

But $L \le 2^{mn}$ so by Fact 1,

L has at most $\Pi(mn)$ prime divisors

Since p is chosen at *random* from $\Pi(M)$ primes,

and only $\Pi(mn)$ give false matches,

$$\text{Prob(false match)} \le \frac{\Pi(mn)}{\Pi(M)} \le \frac{2.511}{n}$$

by Fact 2.

Updating Modular Fingerprints

pattern $X = X_1 \ldots X_m$

text $Y = Y_1 \ldots Y_n$

$X(r) = X,$ $Y(r) = Y_r\, Y_{r+1} \ldots Y_{r+m-1}$

Since $Y(r+1) = (Y(r) - 2^{m-1}\, Y_r) \cdot 2 + Y_{r+m}$

update formula:

$a_p(r+1) = (a_p(r) + a_p(r) + \xi\, Y_r + Y_{r+m}) \bmod p$

where $\xi = -2^m \bmod p$

gives $\Phi_p(Y(r+1)) = a_p(r+1)$ from $\Phi_p(Y(r)) = a_p(r)$

in O(1) arithmetic steps using

O(log n) bit integers on range $[O, mn^2 - 1]$

Theorem

Total Exp. Time for finding a match is O(n)

proof

Expected Time is:

$$O(n) + nm\,\mathrm{Prob}(\text{false match}) \le O(n) + nm\, O\left(\frac{1}{n}\right)$$

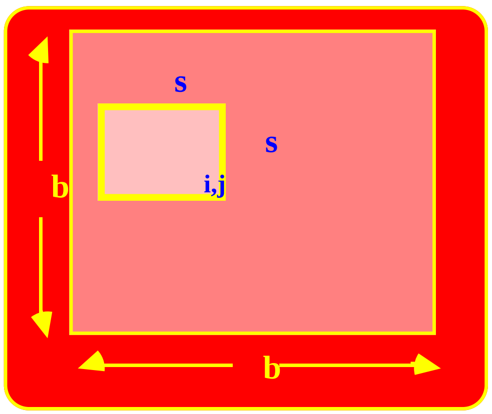
$$\le O(n)$$

2D Randomized Pattern Matching

input pattern $X = (X_{ij})$ is $s \times s$ boul. array

text $Y = (Y_{ij})$ is $b \times b$ boul. array

text window $Y(\langle i,j \rangle)$ = concatination of rows of $s \times s$ subarray of Y with $\langle i,j \rangle$ in lower right corner



Index i Update

fingerprint $a_p(\langle i+1, j \rangle) = a_p(\langle i,j \rangle) +$

$$\left[\left(a_p(\langle i,j \rangle) - \lambda \cdot Y_{i-s+1,j} + \cdot Y_{i+1,j}\right)\right] \bmod p$$

where $\lambda = 2^s \bmod p$

Index j Update

$a_p(\langle i,j+1 \rangle) = a_p(\langle i,j \rangle)\ 2^{-s} +$

$$\left[Y_{i,j+1} \cdot 2 + Y_{2,j+1} \cdot 2^2 + \ldots Y_{s,j+1} \cdot 2^s\right] \theta -$$

$$\left[Y_{i,1} \cdot 2 + Y_{2,1} \cdot 2^2 + \ldots Y_{s,1} \cdot 2^s\right] \bmod p$$

where $\theta = -2^{s(s-1)} \bmod p$

Unimodular Matrices as Fingerprints

Definition

homomorphism k from $\{0,1\}^*$ into unimodular matrices with

$$k(\varepsilon) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad k(0) = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}, \quad k(1) = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$$

$$k(X * Y) = k(X) \cdot k(Y)$$

$*$: concatenation; $\cdot$: matrix multiplication

Fact 1' If $X \in \{0,1\}^m$, then each entry of $k(X)$ is $\leq F_m$ = mth Fibonacci

(where $F_0 = F_1 = 1$, and $F_m = F_{m-1} + F_{m-2}$ for $m \geq 2$)

Fact 2' $\log F_m \sim .694m$

Suppose the unimodular fingerprint fn Φ_p is used with random $p \in S = \{p | p \leq M \text{ is prime}\}$ where $M = mn^2$.

Theorem

The Random Pattern Matching Algorithm has probability $\leq \frac{6.971}{n}$ of false match.

proof

A false match occurs if $\exists i,j \;\varepsilon\; \{1,2\}$

$\Phi_p(X(r))_{i,j} = \Phi_p(Y(r))_{i,j}$ **but** $k(X(r))_{i,j} \neq k(Y(r))_{i,j}$

iff $p \mid L'$ **where** $L' = \prod_{\substack{r\,\varepsilon R \\ i,j\,\varepsilon\{1,2\}}} |k(X(r))_{i,j} - k(Y(r))_{i,j}|$

But $L' \leq (F_m)^{4n} \leq 2^{\lceil 4n \log F_m \rceil}$

By fact 1, the number of primes that divide L' is at most $\Pi\left(\lceil 4n \log F_m \rceil\right)$

The ***probability of a false match*** **is**

$$\frac{\Pi\left(\lceil 4n \log F_m \rceil\right)}{\Pi(M)} \leq \frac{6.971}{n} \quad \text{since by Fact 2'}$$

$\log(F_m) \sim .694m$

Updating Using Unimodular Matrices

1D string matching: $a_p(r) = \Phi_p(Y(r))$

$a_p(r+1) = K_p(Y_r)^{-1}\, a_p(r)\; H(Y_{r+m}) \mod p$

where $K_p(0)^{-1} = \begin{pmatrix} 1 & 0 \\ p-1 & 1 \end{pmatrix}$

and $K_p(1)^{-1} = \begin{pmatrix} 1 & p-1 \\ 0 & 1 \end{pmatrix}$

(can also extend to 2D string matching)

Simplifications of Modular Fingerprinting

S = {p | p≤ M and p prime or pseudoprime}

p is *pseudoprime* if $2^{p-1} \equiv 1 \bmod p$ but p *not* prime

\# *pseudoprimes* is ≤ $Me^{-c(\log m \log\log m)^{\frac{1}{2}}}$

(2) S = {p | p ≤ M and p is M-fat}

p is *M-fat* if p ≤ M and p has prime divisor > $\sqrt{M}$

fact |S| ~ M(ln2 + 0(1))

A false match occurs when p|L

where L = $\prod_{X(r) \neq Y(r)} (H(X(r)) - H(Y(r)))$

Bound L < 2^{nm}

Let N be the number of M-fat integers dividing L

Then $(\sqrt{M})^{\frac{N}{\sqrt{M}}} < L < 2^{nm}$ so

$$N \leq \frac{\ln 2}{2} \frac{n^4}{(\log n)^2}$$

The *prob of false match* is:

$$\frac{N}{|S|} < .5 \quad \text{if} \quad M = \frac{n^4}{(\ln n)^2}$$

(3) $S = \{p \mid p \leq M\}$ with some M

idea use *new* p when get false match

expected time

$$cn\left(.5 + (.5)^2 + (.5)^3 + \ldots\right) \leq O(n)$$

Fingerprinting by Random Polynomials

Galois Field

$GF(2^k) = \{b_1 \ldots b_k \mid b_i, \ldots b_k \in \{0,1\}\}$

$Z_2[t]$ = polynomials of form

$p(t) = t^k + a_{k-1} \cdot t^{k-1} + \ldots + a_0$ where

$a_{k-1}, a_{k-2}, \ldots, a_0 \in \{0,1\}$

$p(x)$ irreducible if can't be factored

Lemma

If k is prime, the number of irreducible polynomials of degree k in $Z_2[t]$

is $\dfrac{(2^k - 2)}{k}$

Fingerprint fn $\Phi_p(x) = x_1 t^{m-1} + \ldots + x_m \bmod p(t)$

(residue comp can be done efficiently)

Theorem

If use random Fingerprint fn with degree $k > \log(nm\ \varepsilon^{-1})$, then prob of false match is $< \varepsilon$.

proof

use usual argument and above Lemma

Open Problems

(1) Are there deterministic methods for Fingerprinting?

(2) What are optimal trade offs for prob of error and size of S for randomized Fingerprinting?

Application of Fingerprinting to Computer Security

for fixed random p,

idea store $\Phi_p(F_1), \ldots, \Phi_p(F_k)$

fingerprints of files $F_1, \ldots, F_k$

Security:

only operator knows p, so if

any file F_i modified, to F_i'

then with high likelihood

$$\Phi_p(F_i') \neq \Phi_p(F_i)$$

$\Rightarrow$ can build a secure operating system

from this idea!