6.006 Rec. 9/30/09

PLAN

1. ROLLING HASH (RECAP)
2. PROBABILISTIC TOOLS
   - UNION BOUND
   - MARKOV'S INEQUALITY
   - LINEARITY OF EXPECTATION
   - AMPLIFICATION
3. BIRTHDAY PARADOX AND THE REVERSE PROBLEM
4. DNA PROBLEM
5. PATTERN MATCHING

ROLLING HASH

$S[1 \ldots n]$

$$h(S[i:i+l]) = \left(\underbrace{S[i]\cdot b^{l-1}} + S[i+1]\cdot b^{l-2} + \cdots + S[i+l-1]\right) \bmod p$$

$$h(S[i+1:i+l+1]) = \left(\left(\quad - \quad\right)\cdot b + S[i+l]\right) \bmod p$$

Can compute $\left(\begin{matrix} b^{l-1} \\ \bmod p \end{matrix}\right)$ in $O(l)$ time

Then $h(S[i+1:i+l+1])$ from $h(S[i:i+l])$ in $O(1)$ time

(2) TOOLS

UNION BOUND

$B_1, \dots, B_k$ - bad events

$$\Pr[\text{any bad event}] \leq \sum_{i=1}^{k} \Pr[B_i]$$

MARKOV'S INEQUALITY

$$X \geq 0$$

random variable

a - any positive real

$$E[X] \geq a \cdot \Pr[X \geq a]$$

that is

$$\Pr[X \geq a] \leq \frac{E[X]}{a}$$

Example:

expected running time: 10s

probability runs longer than 100s: $\leq \frac{10}{100} = \frac{1}{10}$

LINEARITY OF EXPECTATION

$$E[aX + bY] = a \cdot E[X] + b \cdot E[Y]$$

Example:

run 5 times procedure A, expected time of A: 3s

run 2 times B, : 2s

total expected time: $5 \cdot 3s + 2 \cdot 2s = 19s$

AMPLIFICATION

A = ALGORITHM THAT OUTPUTS YES/NO
CORRECT WITH PROBABILITY $\geq 2/3$

RUN k TIMES AND RETURN MAJORITY

CAN SHOW: PROBABILITY OF ERROR $2^{-\Omega(k)}$

BIRTHDAY PARADOX

WE'RE SELECTING RANDOM NUMBERS FROM $\{1, \ldots, N\}$ WITH REPLACEMENT

HOW MANY SAMPLES TO DRAW THE SAME NUMBER TWICE WITH CONSTANT PROBABILITY?

ANSWER: $\Theta(\sqrt{N})$

WE'LL ONLY SHOW LOWER BOUND

SAY, WE PICK k NUMBERS: $a_1 \; a_2 \ldots a_k$

$$\Pr[\underset{i \neq j}{a_i = a_j}] = 1/N$$

$$\Pr[\text{collision}] \underset{\substack{\uparrow \\ \text{UNION BOUND}}}{\leq} \sum_{i \neq j} \Pr[a_i = a_j] = \cancel{\frac{N(N-1)}{2}} \overbrace{\frac{k(k-1)}{2}}^{\binom{k}{2}} \cdot \frac{1}{N}$$

(4) EXAMPLE 1: WANT $\Pr[\text{collision}] \geq \frac{1}{2}$

THEN $\frac{1}{2} \leq \frac{k(k-1)}{2} \cdot \frac{1}{N}$

$$N \leq k(k-1)$$

SO

$$k^2 \geq N$$

$$k \geq \sqrt{N}$$

EXAMPLE 2: n elements hashed into $\{1 \ldots 100\, n^2\}$
uniformly

$$\Pr[\text{collision}] \leq \frac{n(n-1)}{2} \cdot \frac{1}{100 n^2} = \frac{1}{200}$$

DNA PROBLEM

(uniform hashing assumption but we use rolling hash)

ATTEMPT 1: algorithm from the last lecture: create $N = \Theta(n^2)$-size array

expected number of collisions in each iteration =

$$\leq \binom{n}{2} \cdot \frac{1}{N} = O(1)$$

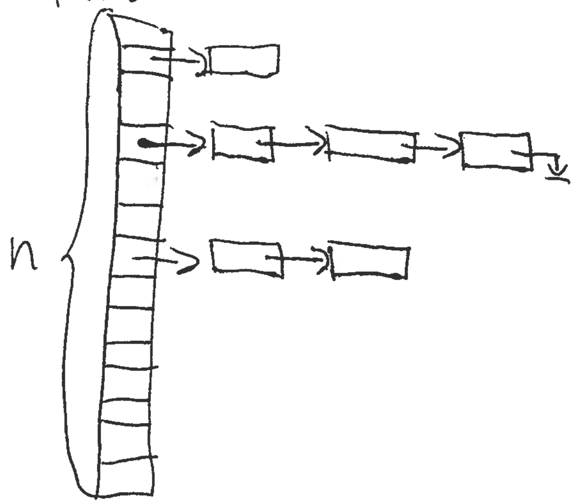
rolling hash: can compute all hash values in $O(n)$ time

expected ~~running~~ time for comparing ~~...~~ strings: $O(1) \cdot O(n) = O($

number of iterations: $O(\log n)$

total expected time: $O(n \log n) + \underbrace{O(n^2)}_{\text{initialization}}$

ATTEMPT 2: HASH INTO A TABLE OF SIZE $\Theta(n)$ BUT COMPARE FIRST SIGNATURES FROM $\{1, \dots, n^2\}$



WHEN INSERTING NEW SUBSTRING, EXPECTED NUMBER OF SIGNATURE COMPARISONS $= O(1)$

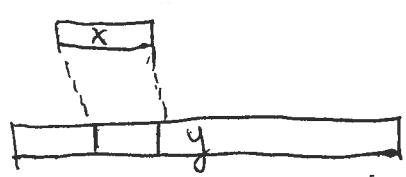
EXPECTED TOTAL NUMBER OF STRING COMPARISONS STILL $O(1)$

EXPECTED RUNNING TIME PER ITERATION: $\underbrace{O(n)}_{\text{ROLLING HASH}} + O(n) \cdot O(1) + O(1) \cdot O(n) = O(n)$

ALL $O(\log n)$ ITERATIONS: $O(n \log n)$

PATTERN MATCHING

FIND AN OCCURENCE OF PATTERN x IN TEXT y



NAIVE SOLUTION: $O(|x| \cdot |y|)$

WANT $O(|y|)$

(6)

1. Compute rolling hash for all subwords of y of length $|x|$
2. Run naive check for subwords y' such that
   $h(y') = h(x)$
   until you find occurence of x

If probability that $h(y') = h(x)$ is $O(1/|x|)$ for any y', then expected running time:

$$O(|y|) + O(|y|) \cdot O(1/|x|) \cdot O(|x|) = O(|y|)$$

rolling hash

This algorithm known as Rabin-Karp algorithm

Our rolling hash invented for this algorithm