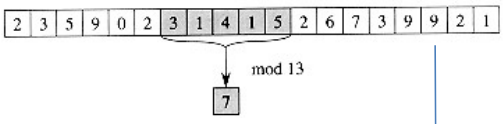
Advanced Algorithm

String Matching

Topics To be Covered

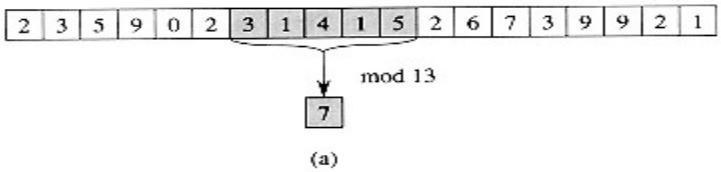
- ✓ String Matching Terminology
- ✓ String Matching Applications
- ✓ Naïve String Matching (Brute-Force Algorithm)
- ✓ Horspool's Algorithm
- ✓ String Matching Using Finite Automata
- ✓ **Rabin-Karp Algorithm** ……and others

Rabin-Karp Algorithm (Part-1)

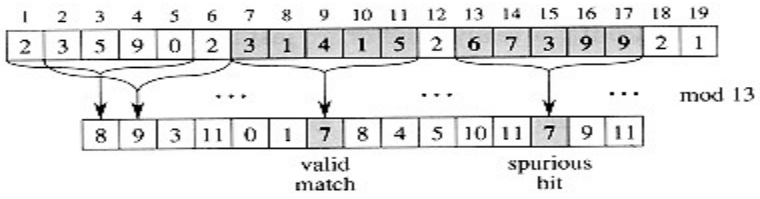



```
RABIN-KARP-MATCHER(T, P, d, q)
1 n ← length[T]
2 m← length[P]
3 h ← d^(m-1) mod q
4 p ← 0
5 t0 ← 0
6 for i ← 1 to m
7         do p ← (dp + P[i]) mod q
8            t0 ← (dt0 + T[i]) mod q
```

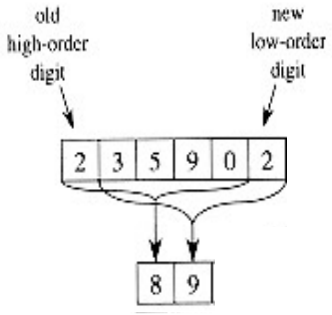
Rabin-Karp Algorithm

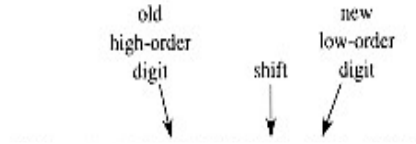


(a)



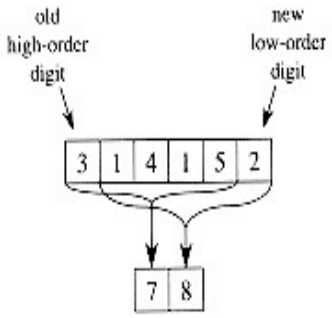
Rabin-Karp Algorithm

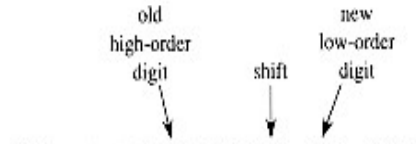




35902 ≡ (23590 – 2 ● 10000) ● 10 + 2 (mod 13)
≡ (8 – 2 ● 3) ● 10 + 2 (mod 13)
≡ (8-6) 10 + 2 (mod 13)
≡ 20 + 2 (mod 13)
≡ 7 + 2 (mod 13)
≡ 9

Rabin-Karp Algorithm





14152 ≡ (31415 – 3·10000)·10 + 2 (mod 13)
≡ (7 – 3·3)·10 + 2 (mod 13)
≡ (7-9) 10 + 2 (mod 13)
≡ -20 + 2 (mod 13)
≡ -7 + 2 (mod 13)
≡ -5
≡ -5 + 13
≡ 8

Rabin-Karp Algorithm (Part-2)

```
9  for s ← 0 to n − m
10      do if p = t_s
11            then if P[1 . . m] = T[s + 1 . . s + m]
12                    then "Pattern occurs with shift" s
13         if s < n − m
14           then t_{s+1} ← (d(t_s − T[s + 1]h) + T[s + m + 1]) mod q
15           if t_{s+1} < 0
16           then t_{s+1} = t_{s+1} + q
```

Rabin-Karp Algorithm

Example-2

- Text: 354861742287
- Pattern: 22
- q=13

Rabin-Karp Algorithm

The running time of RABIN-KARP-MATCHER is $O((n - m + 1)m)$ in the worst case, since (like the naive string-matching algorithm).

The Rabin-Karp algorithm explicitly verifies every valid shift.

If $P = a^m$ and $T = a^n$, then the verifications take time $O((n - m + 1)m)$, since each of the $n - m + 1$ possible shifts is valid.

Rabin-Karp Algorithm

- In many applications, we expect a few valid shifts (perhaps $O(1)$ of them) and some spurious hits, so the expected running time of the algorithm can be calculated as follows:

$O((n - m + 1)m) \approx O(n + m) = O(n) + O(m)$

if we consider only valid shifts & no spurious hits

- The chance that an arbitrary t_s will be equivalent to p, modulo q, can be estimated as $1/q$.
- We can then expect that the number of spurious hits is $O(n/q)$

Rabin-Karp Algorithm

$O((n - m + 1)m) \approx O(\mathrm{n} + \mathrm{m}) \approx O(\mathrm{n}) + O(\mathrm{m})$
$\approx O(n) + O(m(v + n/q))$

Here, v is the number of valid shifts.

if we choose $q >= m$.

$\approx O(n) + O(mv + mn/q))$
$= O(n) + O(mv + mn/m))$
$= O(n) + O(mv + n))$

That is, if the expected number of valid shifts is small ($O(1)$) , then

$= O(n) + O(m+ n))$

we can expect the Rabin-Karp procedure to run in time $O(n + m)$.

Try Yourself

- **Working modulo q = 11, how many spurious hits does the Rabin-Karp matcher encounter in the text T = 3141592653589793 when looking for the pattern P = 26?**