COT 5405: Fall 2006

Lecture 22

Rabin-Karp Algorithm

The Idea

Interpret strings as numbers by interpreting symbols as digits in $\{0, 1, \ldots, |\Sigma|-1\}$.

Example: $\Sigma = \{a, b, c, d, e, f, g, h, i, j\}$. Interpret this as
$\{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$.
So, the string *acdab* is interpreted as *02302*.

Let p be the value of the pattern, P. This can be computed in $O(m)$ time using Horner's rule: $p = P[m] + 10(P[m-1] + 10(P[m-2] + \ldots))$.
t_0 := the value of $T[1 \ldots m]$ can similarly be computed in $O(m)$ time.
Note that $t_{s+1} = 10(t_s - 10^{m-1}T[s+1]) + T[s+m+1]$ can be computed in constant time if we pre-compute10^{m-1} (in $O(m)$ time).

In order to find all matches, we can compare p with t_s for each s. The time complexity is $O(m+n) = O(n)$.

Rabin-Karp Algorithm

The numbers involved in the application of the above idea may be very large. So, we work in modulo $q \geq |\Sigma|$.

Define
- $p' = p\%q$.
- t_0 is defined in a similar manner.

```
Rabin-Karp(T, P)

  • Pre-processing to compute p' and t0
  • for s = 0 to n-m
      o if p' == ts
          ▪ if P[1 … m] = T[s+1 … s+m]
              • Print s
          ▪ if s < n-m
              • ts+1 = [ |Σ|(ts – h T[s+1]) + T[s+m+1] ]%q
              • h = |Σ|^(m-1) % q
```

This takes $\Theta((n-m+1)m)$ time in the worst case.