Rabin Karp

- Submitted by:

  Pujara Chirag S.

Overview Of Presentation

- What it is?
- How it works?
- Reduced Calculation
- Algorithm and Example
- Complexity
- Applications

What Rabin Karp Algorithm is?

- It is String matching algorithm.
- It is another application of Hashing.
- It is widely used for multiple pattern search.

Concept of Rabin Karp Algorithm

- The **Rabin-Karp** string searching algorithm calculates a **hash value** for the pattern, and for each M-character subsequence of text to be compared.
- If the hash values are unequal, the algorithm will calculate the hash value for next M-character sequence.
- If the hash values are equal, the algorithm will compare the pattern and the M-character sequence.
- In this way, there is only one comparison per text subsequence, and character matching is only needed when hash values match.

Some Questions for R.K.

- What is the hash function used to calculate values for character sequences?
- Isn’t it time consuming to hash every one of the M-character sequences in the text body?
- To answer these question we refer to some mathematics.

Some Mathematics for R.K.

- Consider an M-character sequence as an M-digit number in base ***b***, where ***b*** is the number of letters in the alphabet. The text subsequence t[i .. i+M-1] is mapped to the number

  x(i) = ***t***[i]****b***^***M-1***+ ***t***[i+1]****b***^M-2+...+ ***t***[i+M-1]
- Furthermore, given x(i) we can compute x(i+1) for the next subsequence t[i+1 .. i+M] in constant time,as follows:

  x(i+1) = ***t***[i+1]****b***^M-1+ ***t***[i+2]****b***^M-2+...+ ***t***[i+M]

Mathematics Continue

$\boldsymbol{x}(i+1) = \boldsymbol{x}(i)*\boldsymbol{b}$ (Shift left one digit)

$- \boldsymbol{t}[i]*\boldsymbol{b}$^M (Subtract leftmost digit)

$+ \boldsymbol{t}[i+M]$Add new rightmost digit

- In this way, we never explicitly compute a new value. We simply adjust the existing value as we move over one character.
- If M is large, then the resulting value (b^M) will be enormous. For this reason, we hash the value by taking it **mod** a prime number ***q***

Some more Mathematics…

- The **mod** function is particularly useful in this case due to several of its inherent properties:-

  [(x mod q) + (y mod q)] mod q = (x+y) mod q

  (x mod q) mod q = x mod q

- For these reasons:

$\boldsymbol{h}(i) = ((\boldsymbol{t}[i]*\ \boldsymbol{b}\text{^M-1mod }\boldsymbol{q}) + (\boldsymbol{t}[i+1]*\ \boldsymbol{b}\text{^M-2mod }\boldsymbol{q}) + \ldots + (\boldsymbol{t}[i+M-1] \bmod \boldsymbol{q})) \bmod \boldsymbol{q}$

$\boldsymbol{h}(i+1) = (\ \boldsymbol{h}(i)*\ \boldsymbol{b} \bmod \boldsymbol{q}$ (Shift left one digit)

$-\boldsymbol{t}[i]*\ \boldsymbol{b}$^M mod $\boldsymbol{q}$ (Subtract leftmost digit)

$+\boldsymbol{t}[i+M] \bmod \boldsymbol{q}\ )$ (Add new rightmost digit)

mod ***q***

Rabin Karp Algorithm

```
1. n = Length[T],m = Length[P]
2. h = d^m-1 mod q
3. p = 0,to = 0
4. for i= 0 to m
5.     do p = (d*p + P[i]) mod q
6.         to = (d*to + T[i]) mod q
7. For s =0 to n-m
8.     do if p=ts
9.             then if P[1..m] ==T[s+1..s+m]
10.                    then “Pattern founds at shift “ s
11.            if s<n-m
12.            then ts+1=(d(ts-T[s+1])h+T[s+m+1]
```

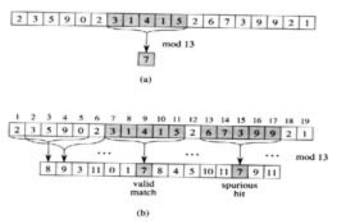
Example of R.K. Algorithm

Calculation for 14152: ts+1=(d(ts-T[s+1]h)+T[s+m+1])mod q

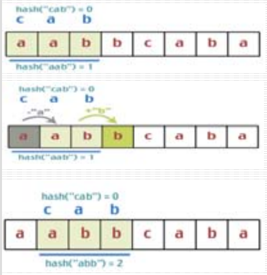
10(31415–3*((10^4)mod13))+2 mod 13

=10(31415- 3*3)+2 mod 13

=8

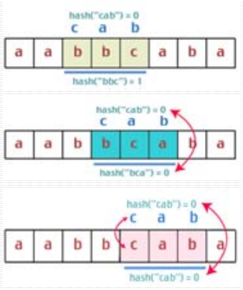


Example

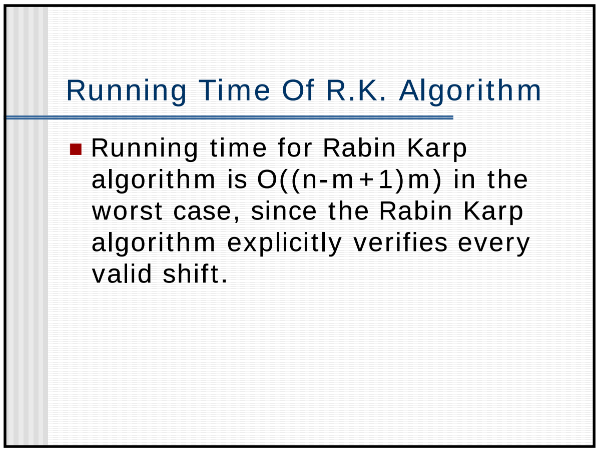


- Let's say our pattern is "cab". And let's say our text string is "aabbcaba".
- For the sake of clarity, we'll use 0 through 26 here to represent letters as opposed to their actual ASCII values.
- For simplicity we will take mod 3.

Example

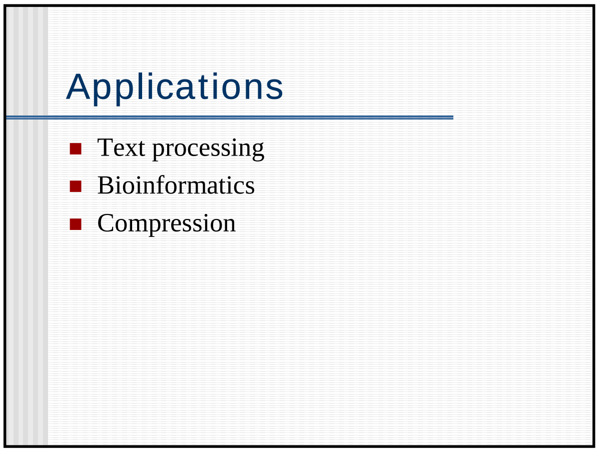


- Here we have collision but it is ignored by code line 8-10



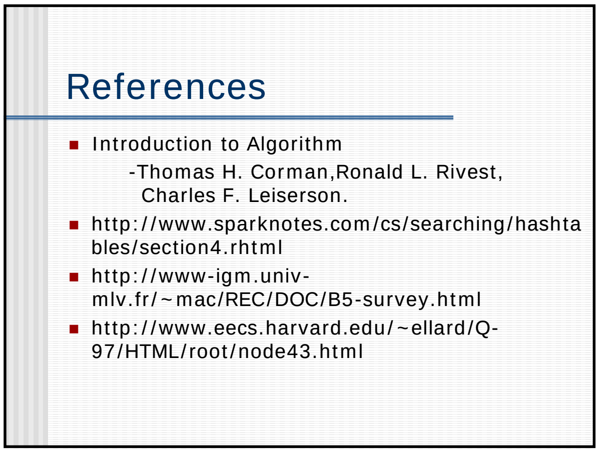
Running Time Of R.K. Algorithm

- Running time for Rabin Karp algorithm is O((n-m+1)m) in the worst case, since the Rabin Karp algorithm explicitly verifies every valid shift.



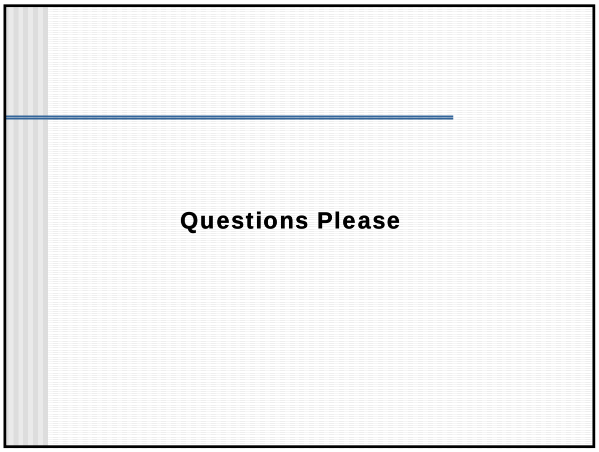
Applications

- Text processing
- Bioinformatics
- Compression



References

- Introduction to Algorithm
  -Thomas H. Corman,Ronald L. Rivest,
  Charles F. Leiserson.
- http://www.sparknotes.com/cs/searching/hashtables/section4.rhtml
- http://www-igm.univ-mlv.fr/~mac/REC/DOC/B5-survey.html
- http://www.eecs.harvard.edu/~ellard/Q-97/HTML/root/node43.html



Questions Please