7. STRING MATCHING

Serge Gaspers
serge.gaspers@unsw.edu.au

Katie Clinch
k.clinch@unsw.edu.au

Course admin: Song Fang
cs3121@cse.unsw.edu.au

School of Computer Science and Engineering, UNSW Sydney

Term 3, 2023

Table of Contents

1. Introduction

2. Hashing

3. Finite Automata
   - How to compute the transition function δ
   - How to solve the substring matching problem in linear time

4. Imperfect matches

5. Puzzle

String matching algorithms

Suppose you have an alphabet consisting of d symbols. Strings are just sequences of these symbols.

Problem

Determine whether a string $B = b_1 b_2 \dots b_m$ (the pattern) appears as a (contiguous) substring of a much longer string $A = a_1 a_2 \dots a_n$ (the text).

String matching problem

Attempt 1

The "naïve" algorithm for string matching is as follows: for every position i in A, check $a_i \ldots a_{i+m-1}$ against B character-by-character.

Even with early exit, this runs in $O(mn)$ time in the worst case.

Question

Can we do better?

Table of Contents

1. Introduction
2. Hashing
3. Finite Automata
    - How to compute the transition function δ
    - How to solve the substring matching problem in linear time
4. Imperfect matches
5. Puzzle

String matching with hashing

- We now show how hashing can be combined with recursion to produce an efficient string matching algorithm.

- We compute a hash value for the string $B = b_1 b_2 \ldots b_m$ in the following way.

- First, map each symbol to a corresponding integer $i \in \{0, 1, 2, \ldots, d-1\}$ so as to identify each string with a sequence of these integers (recall that d is the number of symbols in our alphabet).

- Hereafter, when we refer to an integer a_i or b_i, we really mean the ID of the symbol a_i or b_i.

Polynomial rolling hash

- We can therefore identify B with a sequence of IDs $\langle b_1, b_2, \ldots, b_{m-1}, b_m \rangle$, each between 0 and $d-1$ inclusive. Viewing these IDs as digits in base d, we can construct a corresponding integer

$$\begin{aligned} h(B) &= h(b_1 b_2 \ldots b_{m-1} b_m) \\ &= d^{m-1} \times b_1 + d^{m-2} \times b_2 + \ldots + d \times b_{m-1} + b_m. \end{aligned}$$

- This can be evaluated efficiently using Horner's rule:

$$h(B) = b_m + d \times (b_{m-1} + d \times (b_{m-2} + \ldots + d \times (b_2 + d \times b_1) \ldots)),$$

requiring only $m-1$ additions and $m-1$ multiplications.

- Next we choose a large prime number p and define the hash value of B as $H(B) = h(B) \bmod p$.

Polynomial rolling hash

- Recall that $A = a_1 a_2 a_3 \ldots\ldots a_s a_{s+1} \ldots a_{s+m-1} \ldots\ldots a_n$ where $n \gg m$.

- We want to find efficiently all s such that the string of length m of the form $a_s a_{s+1} \ldots a_{s+m-1}$ and string $b_1 b_2 \ldots b_m$ are equal.

- For each contiguous substring $A_s = a_s a_{s+1} \ldots a_{s+m-1}$ of string A we also compute its hash value as

  $$H(A_s) = d^{m-1} a_s + d^{m-2} a_{s+1} + \ldots + d^1 a_{s+m-2} + a_{s+m-1} \bmod p.$$

Polynomial rolling hash

- We can now compare the hash values $H(B)$ and $H(A_s)$.
  - If they disagree, we know that there is no match.
  - If they agree, there might be a match. To confirm this, we'll check it character-by-character.
- There are $O(n)$ substrings A_s to check. If we compute each hash value $H(A_s)$ in $O(m)$ time, this is no better than the naïve algorithm.
- This is where recursion comes into play: we do not have compute the hash value $H(A_{s+1})$ "from scratch". Instead, we can compute it efficiently from the previous hash value $H(A_s)$.

Polynomial rolling hash

- Recall that $A_s = a_s a_{s+1} \ldots a_{s+m-1}$, so

$$H(A_s) = d^{m-1} a_s + d^{m-2} a_{s+1} + \ldots + a_{s+m-1} \bmod p.$$

- Multiplying both sides by d gives

$$\begin{aligned} d \cdot H(A_s) &= d^m a_s + d^{m-1} a_{s+1} + \ldots + d^1 a_{s+m-1} \\ &= d^m a_s + \left(d^{m-1} a_{s+1} + \ldots + d^1 a_{s+m-1} + a_{s+m}\right) - a_{s+m} \\ &= d^m a_s + H(A_{s+1}) - a_{s+m} \bmod p, \end{aligned}$$

since $A_{s+1} = a_{s+1} \ldots a_{s+m-1} a_{s+m}$.

Polynomial rolling hash

- Rearranging the last equation, we have

$$H(A_{s+1}) = d \cdot H(A_s) - d^m a_s + a_{s+m} \bmod p.$$

- Thus we can compute $H(A_{s+1})$ from $H(A_s)$ in constant time.

Note

When implementing this, we would like to choose large primes to reduce the likelihood of hash collisions. However, this is constrained by the intermediate values in the calculation above, which are up to $d \times p$, and must not overflow a register.

Rabin-Karp Algorithm

Algorithm

1. First compute $H(B)$ and $H(A_1)$ in $O(m)$ time using Horner's rule.

2. Then compute the $O(n)$ subsequent values of $H(A_s)$ each in constant time using the recurrence above.

3. Compare each $H(A_s)$ with $H(B)$, and if they are equal then confirm the potential match by checking the strings A_s and B character-by-character.

Rabin-Karp Algorithm

- Since p was chosen large, the false positives (where $H(A_s) = H(B)$ but $A_s \neq B$) are very unlikely, which makes the algorithm run fast (linear) in the average case.

- However, as always when we use hashing, we cannot achieve useful bounds for the worst case performance.

- So we now look for algorithms whose worst case performance is guaranteed to be linear.

Table of Contents

1. Introduction

2. Hashing

3. Finite Automata
    - How to compute the transition function δ
    - How to solve the substring matching problem in linear time

4. Imperfect matches

5. Puzzle

String matching with finite automata

- A string matching finite automaton for a pattern B of length m has:
    - $m+1$ many states $0, 1, \dots, m$, which correspond to the number of characters matched thus far, and
    - a transition function $\delta(k, a)$, where $0 \leq k \leq m$ and a is a character in your alphabet.
- Suppose that the last k characters of the text A match the first k characters of the pattern B, and that a is the next character in the text. Then $\delta(k, a)$ is the new state after character a is read, i.e. the largest k' so that the last k' characters of A (ending at the new character a) match the first k' characters of B.

String matching with finite automata

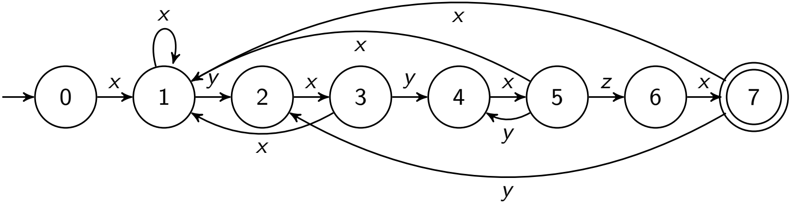
As an example, consider the string $B = xyxyxzx$. The table defining $\delta(k, a)$ would then be as follows:

k	matched	x	y	z
0		**1**	0	0
1	x	1	**2**	0
2	xy	**3**	0	0
3	xyx	1	**4**	0
4	$xyxy$	**5**	0	0
5	$xyxyx$	1	4	**6**
6	$xyxyxz$	**7**	0	0
7	$xyxyxzx$	1	2	0

String matching with finite automata

This table can be visualised as a state transition diagram. From each state k, we draw an arrow to $\delta(k, a)$ for each character a that could be encountered next. Arrows pointing to state 0 have been omitted for clarity.

$B = xyxyxzx$



String matching with finite automata

- **How do we build the transition diagram?**

- **How do we build the transition table?**

These questions are equivalent to asking:

- **How do we compute the transition function δ?**

In the remainder of the lecture we will:

1. Explain how to compute the transition function δ.
2. Use this to construct an algorithm that solves our original substring matching problem (see slide 3) in linear time.

Table of Contents

1. Introduction

2. Hashing

3. Finite Automata
    - How to compute the transition function δ
    - How to solve the substring matching problem in linear time

4. Imperfect matches

5. Puzzle

Computing the transition function δ

Observation. We constructed our example transition table and transition diagram by comparing suffixes of our currently checked string against prefixes of B to find the longest match.

We first need to formalise these ideas.

- Let $B_k = b_1 \dots b_k$ denote a prefix of length k of the string B.
- Being at state k means that so far we have matched the prefix B_k.
- If we now see an input character a, then $\delta(k, a)$ is the largest ℓ such that the prefix B_ℓ of string B is a suffix of the string $B_k a$.

Computing the transition function δ

Question

How do we find $\delta(k, a)$, the largest ℓ such that B_ℓ is a suffix of $B_k a$?

- If a happens to match with b_{k+1}, i.e. $B_k a = B_{k+1}$, then $\ell = k + 1$ and so $\delta(k, a) = k + 1$.
- But what if $a \neq b_{k+1}$? Then we can't extend our match from length k to $k + 1$.
- We'd like to extend some shorter match instead.
- Any other candidates to be extended must be both a proper suffix of B_k and a prefix of B.

The failure function

Definition

Let $\pi(k)$ be the largest integer ℓ such that the prefix B_ℓ of B is a proper suffix of B_k.

B_ℓ is therefore the longest substring which appears at both the start and the end of B_k (allowing partial but not total overlap).

We will refer to π as the failure function.

Note that some texts refer to π as the prefix function.

The failure function

The transition function $\delta(k, a)$ is closely related to the failure function $\pi(k)$:

- If $a = b_{k+1}$, then $\delta(k, a) = k + 1$.
- Otherwise, we look to extend a shorter match. The next candidate has length $\pi(k)$, so we check whether $a = b_{\pi(k)+1}$ and if so we have $\delta(k, a) = \pi(k) + 1$.
- Otherwise, the next candidate has length $\pi(\pi(k))$ (why?) and so on.

Observation

- If we know the failure function $\pi(k)$, then computing the transition function $\delta(k, a)$ is easy!
- In fact, we can describe the algorithm and analysis in terms of π and not use δ at all.
- So how do we compute the failure function $\pi(k)$?

Computing the failure function π

- Our first step is to solve the value of the failure function for each index of the pattern B.

- We will do this recursively using dynamic programming.

- The subproblems of the DP are the $\pi(k)$ values for each $1 \leq k \leq m$, and the base case is $\pi(1) = 0$.

- Suppose we have solved $\pi(1), \ldots, \pi(k)$. How do we find $\pi(k+1)$?

Computing the failure function π

- Recall that $B_{\pi(k)} = b_1 \ldots b_{\pi(k)}$ is the longest prefix of B which is a proper suffix of $B_k = b_1 \ldots b_k$.

- To compute $\pi(k+1)$, we first try to extend this match, by checking whether $b_{k+1} = b_{\pi(k)+1}$.
    - If true, then $\pi(k+1) = \pi(k) + 1$.
    - If false, then we cannot extend $B_{\pi(k)}$. What's the next longest prefix of B which is a proper suffix of B_k?
    - It's $B_{\pi(\pi(k))}$! So we check whether $b_{k+1} = b_{\pi(\pi(k))+1}$.
        - If true, then $\pi(k+1) = \pi(\pi(k)) + 1$.
        - If false, check whether $b_{k+1} = b_{\pi(\pi(\pi(k)))+1}$ ...

Computing the failure function π

```
function COMPUTE-FAILURE-FUNCTION($B$)
    $m \leftarrow |B|$
    let $\pi[1..m]$ be a new array, and $\pi[1] \leftarrow 0$
    $\ell \leftarrow 0$
    for $k = 1$ to $m - 1$ do
        while $\ell \geq 0$ do
            if $b_{k+1} = b_{\ell+1}$ then
                $\ell \leftarrow \ell + 1$
                break
            end if
            if $\ell > 0$ then
                $\ell \leftarrow \pi[\ell]$
            else
                break
            end if
        end while
        $\pi[k + 1] \leftarrow \ell$
    end for
    return $\pi$
end function
```

Computing the failure function π

What is the complexity of this algorithm?

- There are $O(m)$ values of k, and for each we might try several values ℓ; is this $O\left(m^2\right)$?

- No! It is actually linear, i.e. $O(m)$.

- Maintain two pointers: the left pointer at $k - \ell$ (the start point of the match we are trying to extend) and the right pointer at $k + 1$.

- After each 'step' of the algorithm (i.e. each comparison between b_{k+1} and $b_{\ell+1}$), exactly one of these two pointers is moved forwards.

- Each can take up to m values, so the total number of steps is $O(m)$.

Table of Contents

1. Introduction

2. Hashing

3. Finite Automata
   - How to compute the transition function δ
   - How to solve the substring matching problem in linear time

4. Imperfect matches

5. Puzzle

The Knuth-Morris-Pratt algorithm

- We can now do our search for the pattern B in the text A.
- Suppose B_s is the longest prefix of B which is a suffix of $A_i = a_1 \ldots a_i$.
- To answer the same question for A_{i+1}, we begin by checking whether $a_{i+1} = b_{s+1}$.
    - If true, then the answer for A_{i+1} is $s+1$
    - If false, check whether $a_{i+1} = b_{\pi(s)+1} \ldots$
- If the answer for any A_i is m, we have a match!
    - Reset to state $\pi(m)$ to detect any overlapping full matches.
- By the same two pointer argument, the time complexity is $O(n)$.

The Knuth-Morris-Pratt algorithm

```
function KMP-Matcher($A, B$)
    $n \leftarrow |A|$
    $m \leftarrow |B|$
    $\pi \leftarrow$ Compute-Failure-Function($B$)
    $s \leftarrow 0$
    for $i = 0$ to $n - 1$ do
        while $s \geq 0$ do
            if $a_{i+1} = b_{s+1}$ then
                $s \leftarrow s + 1$
                break
            end if
```

The Knuth-Morris-Pratt algorithm

```
            **if** $s > 0$ **then**
                $s \leftarrow \pi[s]$
            **else**
                **break**
            **end if**
        **end while**
        **if** $s = m$ **then**
            **print** match found, ending at index $i + 1$
            $s \leftarrow \pi[m]$
        **end if**
    **end for**
**end function**
```

Table of Contents

1. Introduction
2. Hashing
3. Finite Automata
    - How to compute the transition function δ
    - How to solve the substring matching problem in linear time
4. Imperfect matches
5. Puzzle

Looking for imperfect matches

Sometimes we are not interested in finding just the perfect matches, but also in matches that might have a few errors, such as a few insertions, deletions and replacements.

Problem

Instance: a text

$$A = a_1 a_2 \ldots \ldots a_s a_{s+1} \ldots a_{s+m-1} \ldots \ldots a_n,$$

a pattern $B = b_1 b_2 \ldots b_m$ where $m \ll n$, and an integer $k \ll m$.

Task: find all matches for B in A which have up to k errors.

Looking for imperfect matches

Solution Outline

Split B into $k+1$ substrings of (approximately) equal length. Within any match in A with at most k errors, at least one of these $k+1$ parts of B must be matched perfectly.

For each of the $k+1$ parts of B, we use the Knuth-Morris-Pratt algorithm to find all perfect matches in A.

Then for each of these matches, we test by brute force whether the remaining parts of B match sufficiently with the appropriate parts of A.

Table of Contents

1. Introduction

2. Hashing

3. Finite Automata
   - How to compute the transition function δ
   - How to solve the substring matching problem in linear time

4. Imperfect matches

5. Puzzle

Puzzle

On a rectangular table there are 25 round coins of equal size, and no two of these coins overlap. You observe that in current arrangement, it is not possible to add another coin without overlapping any of the existing coins and without the coin falling off the table (for a coin to stay on the table its centre must be within the table).

Show that it is possible to completely cover the table with 100 coins (allowing overlaps).