

Pemanfaatan Algoritma Boyer Moore dalam Penyaringan Teks Halaman Website Sederhana

Rheno Manggala Budiasa and 13506119¹
Program Studi Teknik Informatika
Sekolah Teknik Elektro dan Informatika
Institut Teknologi Bandung, Jl. Ganesha 10 Bandung 40132, Indonesia
¹if16119@students.if.itb.ac.id

Abstrak—Aplikasi web mulai banyak digunakan saat ini. Bukan hanya di Indonesia bahkan di dunia. Seiring berkembangnya Teknologi Informasi, pertumbuhan *website* dari tahun ke tahun semakin banyak. Istilah *Web 2.0* (bahkan sekarang *Web 3.0*) mungkin sudah akrab di telinga kita. Istilah yang sudah akrab sejak 5 tahun belakangan ini sebenarnya adalah sebuah kesepakatan beberapa pengembang web mengenai sebuah standarisasi dari teknologi web. Hal ini penting mengingat, Pengembang *Web* membutuhkan acuan dalam mengembangkan aplikasi mereka. Karena itu standarisasi ini menjadi sangat penting. Banyaknya Halaman *Website* di satu sisi menjadi hal yang baik, mengingat pengguna dapat melihat berbagai informasi yang dibutuhkan. Di sisi lain hal ini bisa menjadi bumerang apabila kita menyadari bahwa tidak semua halaman *Web* baik, terutama untuk anak-anak. Seperti kita tahu juga bahwa berbagai informasi bisa kita dapatkan di internet saat ini, mulai dari konten yang baik hingga buruk. Karena itu dibutuhkan suatu program atau metode yang dapat membatasi akses terhadap sebuah halaman *Website*.

Kata Kunci—Aplikasi, Web, *Web 2.0*, kesepakatan, pengembang, standarisasi, Halaman, pengguna, konten, membatasi, akses.

1. I. Pendahuluan

Pembatasan terhadap konten web sebenarnya merupakan masalah yang sangat serius sampai saat ini. Berbagai metode sudah digunakan untuk menyelesaikan masalah ini. Mulai dari level Aplikasi sampai Jaringan. Khusus untuk level Aplikasi ada beberapa metode yang dapat digunakan salah satunya adalah penyaringan terhadap isi dari suatu halaman *Web*. Banyak algoritma yang dapat kita manfaatkan untuk kasus ini. Beberapa algoritma yang sering ada antara lain *Brute Force*, Boyer-Moore, KMP, dan lain-lain.

Metode dari penyaringan *Website* biasanya bermacam-macam namun yang sering digunakan adalah isi teks dari *Tag HTML*. Melalui metode ini setiap halaman *Web* yang mengandung kata-kata yang tidak dikehendaki akan otomatis *block* atau tidak dapat diakses.

Program penyaringan ini pada kenyataannya bisa diletakkan pada suatu *Proxy*. Tidak hanya itu program penyaringan tersebut juga dapat dibuat pada browser dari sisi *client*.

2.

3. II. Teori

Dasar dari penyelesaian persoalan ini adalah *Pattern Matching*. Boyer-Moore merupakan salah satu Algoritma *Pattern Matching* yang cukup terkenal. Algoritma ini menggunakan beberapa kasus pengecekan Teks (input karakter yang akan dibaca) dengan *Pattern* (pola yang akan di saring).

Algoritma *Pattern Matching* Boyer-Moore ini berbasis pada 2 metode yaitu :

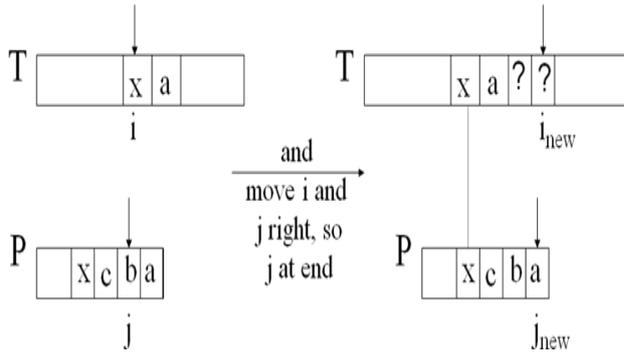
1. The Looking-Glass Technique
 - Mencari Suatu kecocokan String pada Teks dengan pola yang akan dicari dengan cara memindahkan atau menggesernya sampai Teks string selesai.
2. The Character-Jump Technique
 - Mencari karakter yang sesuai dan cara penggesaran sebuah karakter perbandingan terakhir.

Beberapa kasus yang ada pada algoritma ini antara lain :

1. Jika suatu karakter Pola (P) mengandung karakter x dimana x adalah anggota dari Teks yang telah dibandingkan. Maka Perbandingan Karakter Selanjutnya dimulai karakter P yang sama dengan

Misal :

T = ..xa..??
P = xcba

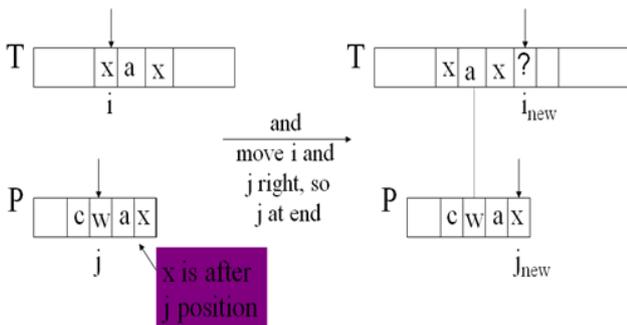


Gambar 1
Kasus Pertama

2. Jika Perbandingan karakter terakhir pada suatu pola sama dengan teks adalah sama. Maka pergeseran karakter selanjutnya bergeser satu kali.

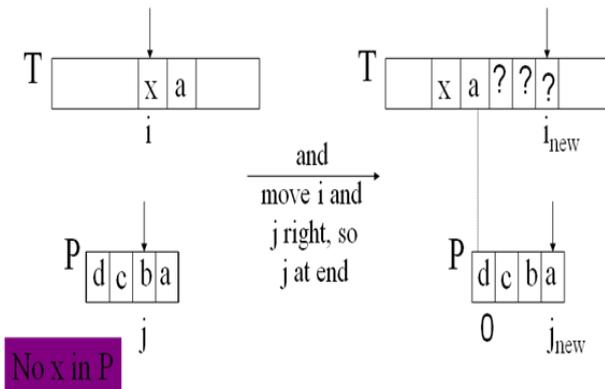
Misal :

T = ..xax..??
P = cwax



Gambar 2
Kasus Kedua

3. Jika Kondisi 1 dan 2 tidak terpenuhi maka perbandingan karakter pada P dimulai dari Karakter perbandingan akhir T dengan Karakter pertama P.



No x in P

Gambar 3 Kasus Ketiga

Contoh Algoritma *Pattern-Matching* Boyer-Moore.

T:

P:

Gambar 4 Pattern-Matching Boyer-Moore

Kasus terburuk terjadi jika semua Teks (T) merupakan sebuah kata yang mempunyai karakter yang sama dan sangat panjang, dengan Pola (P) dengan karakter awal yang berbeda dengan semua karakter yang ada pada (T) :

T:

P:

Gambar 5 Kasus Terburuk

Untuk kasus Rata-rata Kompleksitas waktunya adalah:

$O(n/m)$

Kompleksitas Waktu pada kasus terburuk adalah :

$O(nm)$

Dibandingkan dengan Algoritma *Brute Force*, Boyer-Moore sebenarnya sudah jauh lebih baik. Kalau kita bandingkan secara kompleksitas waktu, Brute Force memiliki kasus $O(nm)$. Karena itu kasus Rata-rata Boyer-Moore cukup menjadi solusi yang optimal.

Hyper Text Markup Level (HTML) merupakan bahasa pemrograman yang berjalan di sebuah Web. HTML memerlukan browser untuk mengubah setiap kodenya untuk menjadi sebuah halaman. HTML terdiri dari beberapa struktur bersarang yang masing-masing mempunyai fungsi yang berbeda-beda. Struktur ini disebut Tag.

Ada berbagai macam tag pada HTML, namun setidaknya sebuah situs memiliki paling tidak beberapa struktur standar :

```
<html>  
<head>  
<title>Tag Standar</title>
```

```

</head>
<body>
  <p>Ini Tag Standar</p>
</body>
</html>

```

Dalam melihat isi halaman web biasanya tag `<body></body>` digunakan sebagai konten pada suatu website. Sedangkan `<title></title>` digunakan sebagai judul dari sebuah konten web. Karena itu banyak Mesin Pencari (seperti Google dan Yahoo) menggunakan Tag-tag ini untuk mencari isi dari suatu halaman web.

Preprocessor Hypertext Preprocessor (PHP) merupakan bahasa pemrograman tingkat tinggi yang berjalan pada suatu *Web Server*. Karena sifatnya yang bisa berjalan pada Server maka bahasa ini sering dijadikan standar baku bagi para pengembang web sebagai bahasa standar dalam pembuatan suatu web bersama dengan HTML. PHP juga bisa digabungkan dengan HTML. Karena itu kedua bahasa ini biasanya digunakan sebagai bersamaan. HTML biasanya digunakan untuk menampilkan data baik itu berupa teks, gambar, video dan lain-lain. Sedangkan PHP biasanya digunakan sebagai koneksi antara server dengan client. Kode PHP tidak dapat terlihat di sisi client. Karena itu dari segi keamanan PHP cukup handal karena tidak dapat dirubah atau dibaca oleh client.

Sama seperti HTML, PHP mempunyai struktur yang standar :

```

<?php
echo"
<html>
  <head>
    <title>Tag Standar</title>
  </head>
  <body>
    <p>Ini Tag Standar</p>
  </body>
</html>";
?>

```

Kode di atas menampilkan kode HTML pada PHP. Sebaliknya PHP juga memungkinkan untuk menampilkan sebaliknya yaitu menampilkan PHP di HTML :

```

<html>
  <head>
    <?php echo"<title>Tag Standar</title>"; ?>
  </head>
  <body>
    <?php echo"<p>Ini Tag Standar</p>"; ?>
  </body>
</html>

```

PHP banyak digunakan karena kemampuannya yang sangat banyak. PHP dapat digunakan sebagai penghubung ke Database.

4. III. Program

1. A. Algoritma

Strategi pertama adalah membuat Algoritma pemrograman. Hal ini dibutuhkan pertama kalo supaya kita dapat lebih mudah nantinya dalam membuat kode.

```

function buildLastFunction(input pattern :
string):integer

Deklarasi :
Last : array[1..128] of integer
i: integer

Algoritma :
for i = 0 to 128 do
  last[i] = -1
  for i = 0 to pattern.length do
    last[pattern.charAt(i)] = i
-> last
function BMmatch(input text : string, pattern :
string):integer

Deklarasi :
last : array[1..128] of integer
i: integer
n : integer
m : integer
j: integer

Algoritma :
last = buildLastFunction(pattern)
n = text.length
m = pattern.length
i = m-1
-> -1
j=m-1
do
  if pattern.charAt(i) = text.charAt(i)
    if j = 0
      ->I
    else
      i--; j--
    else
      i = i+m - Min(j, 1+last[charAt(i)])
      j = m-1
While i <= -1

-> -1

```

Gambar 6
Algoritma Boyer-Moore

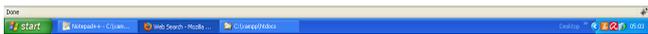
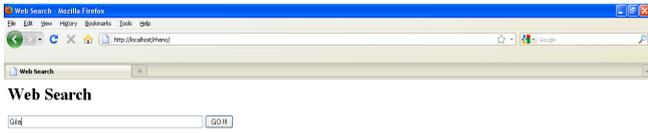
B. Struktur Program

Program terdiri dari 3 buah, Pertama, Program untuk menerima inputan dari user. Program ini dapat digunakan untuk mencari beberapa halaman web. Halaman yang dimaksud adalah halaman HTML-nya.

Program kedua merupakan inti dari algoritma boyer-moore. Program ini akan membaca setiap tulisan yang ada pada tag <title> dan <body> pada kode HTML.

Program ketiga merupakan program yang berfungsi untuk menampilkan hasil. Jika Hasil dari halaman suatu web memuat kata yang tidak diinginkan, maka halaman ini akan menampilkan pesan halaman ini mengandung kata yang tidak baik. Namun jika tidak halaman ini akan memuat semua konten web.

Ilustrasi program dapat dilihat di bawah ini :



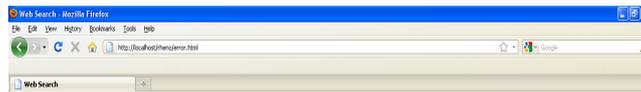
Gambar 7
Program Pertama

Program kedua algoritma (PHP):

```
public function bmMatch($text, $pattern)
{
    $last = array();
    $n = count(text );
    $m = count(pattern);
    $i = $m-1;

    if ($i > $n-1)
        return -1;
    $j = $m-1;
    do {
        if ($pattern[$j] == $text[$i])
            if ($j == 0)
                return $i; // match
            else {
                $i--;
                $j--;
            }
        else {
            $lo = last[text[$i]];
            $i = $i + $m - Min(j, 1 + $lo);
            $j = $m-1;
        }
    } while ($i <= $n-1);

    return -1; }
```



Web site ini Mengandung Kata yang tidak baik

[Link](#)



Gambar 8
Program Ketiga Menampilkan Pesan



Gambar 9
Langsung redirect ke halaman Web

2. C. Abbreviations and Acronyms

Define abbreviations and acronyms the first time they are used in the text, even after they have already been defined in the abstract. Abbreviations that incorporate periods should not have spaces: write “C.N.R.S.,” not “C. N. R. S.” Do not use abbreviations in the title unless they are unavoidable.

5. IV. Kesimpulan

- Algoritma Boyer-Moore salah satu optimalisasi dari pattern Matching
- Algoritma ini bisa digunakan untuk teknologi mesin pencari
- Untuk Kasus terburuk algoritma ini tidak mangkus. Karena Kompleksitas waktunya sama dengan *Brute Force*

6. Referensi

1. Rinaldi Munir, *Bahan Kuliah IF2251 Strategi Algoritma*, 2008

2. www.w3schools.com
3. Dr. Andrew Davison, *Pattern Matching*, WiG Lab (teachers room), CoE, 2006-2007

7.

8. PERNYATAAN

Dengan ini saya menyatakan bahwa makalah yang saya tulis ini adalah tulisan saya sendiri, bukan saduran, atau terjemahan dari makalah orang lain, dan bukan plagiasi.

Bandung, 9 Desember 2010

ttd

Rhen Manggala Budiasa 13506119