Text Algorithms

Jaak Vilo
2016 fall
• Exact matching of one pattern(string)
• Exact matching of multiple patterns
• Suffix trie and tree indexes
  – Applications
• Suffix arrays
• Inverted index
• Approximate matching
Algorithms

One-pattern
- Brute force
- Knuth-Morris-Pratt
- Karp-Rabin
- Shift-OR, Shift-AND
- Boyer-Moore
- Factor searches
- Regular expressions(?)
- Weight matrices(?)

Multi-pattern
- Aho Corasick
- Commentz-Walter

Indexing
- Trie (and suffix trie)
- Suffix tree
Exact pattern matching

- \( S = s_1 s_2 \ldots s_n \) (text) \(|S| = n\) (length)

- \( P = p_1 p_2 \ldots p_m \) (pattern) \(|P| = m\)

- \( \Sigma \) - alphabet \(|\Sigma| = c\)

- Does \( S \) contain \( P \)?
  - Does \( S = S' P S'' \) for some strings \( S' \) ja \( S'' \)?
  - Usually \( m \ll n \) and \( n \) can be (very) large
Find occurrences in text
Animations


• EXACT STRING MATCHING ALGORITHMS Animation in Java

• Christian Charras - Thierry Lecroq
  Laboratoire d'Informatique de Rouen
  Université de Rouen
  Faculté des Sciences et des Techniques
  76821 Mont-Saint-Aignan Cedex
  FRANCE

• e-mails: {Christian.Charras, Thierry.Lecroq}@laposte.net
Brute force: BAB in text?

Brute Force

Identify the first mismatch!

Question:

- Problems of this method? 😞
- Ideas to improve the search? 😊
Brute force

Algorithm Naive

**Input:** Text $S[1..n]$ and pattern $P[1..m]$

**Output:** All positions $i$, where $P$ occurs in $S$

```plaintext
for ( i=1 ; i <= n-m+1 ; i++ )
  for ( j=1 ; j <= m ; j++ )
    if ( $S[i+j-1] != P[j]$ ) break;
  if ( j > m ) print i;
```

```
attempt 1:
gcatgcgagaggtacgtacg
GCAG....

attempt 2:
gcatcgagaggtacgtacg
g........

attempt 3:
gcatcgagaggtacgtacg
g........

attempt 4:
gcatcgagaggtacgtacg
g........

attempt 5:
gcatcgagaggtacgtacg
g........

attempt 6:
gcatcGAGAGGtacgtacg
GCAAGAG

attempt 7:
gcatcGAGAGGtacgtacg
g........
```
Brute force or NaiveSearch

1 function NaiveSearch(string s[1..n], string sub[1..m])
2 for i from 1 to n-m+1
3 for j from 1 to m
4 if s[i+j-1] ≠ sub[j]
5 jump to next iteration of outer loop
6 return i
7 return not found
C code

```c
int bf_2( char* pat, char* text, int n ) /* n = textlen */
{
    int m, i, j;
    int count = 0;
    m = strlen(pat);

    for ( i=0 ; i + m <= n ; i++) {
        for( j=0; j < m && pat[j] == text[i+j] ; j++) ;

        if( j == m )
            count++ ;

    }

    return(count);
}
```
```c
int bf_1( char* pat, char* text )
{
    int m ;
    int count = 0 ;
    char *tp;

    m = strlen(pat);
    tp=text ;

    for( ; *tp ; tp++ ) {
        if( strncmp( pat, tp, m ) == 0 ) {
            count++ ;
        }
    }

    return( count );
}
```
Main problem of Naive

- For the next possible location of $P$, check again the same positions of $S$
Goals

• Make sure only a constant nr of comparisons/operations is made for each position in S
  – Move (only) from left to right in S

  – How?
  – After a test of S[i] <> P[j] what do we now?
Knuth-Morris-Pratt

• Make sure that no comparisons “wasted”

• After such a mismatch we already know exactly the values of green area in S!
Knuth-Morris-Pratt

- Make sure that no comparisons “wasted”

- P – longest suffix of any prefix that is also a prefix of a pattern

- Example: 

  \[ \text{ABCABD} \]

  \[ \underline{AB} \underline{C} \underline{A} \underline{B} \underline{D} \]
Automaton for ABCABD
Automaton for ABCABD

NOT A

Fail links: 0 1 1 1 2 3 1

Pattern: A B C A B D

1 2 3 4 5 6
KMP matching

**Input:** Text $S[1..n]$ and pattern $P[1..m]$

**Output:** First occurrence of $P$ in $S$ (if exists)

```
  i=1; j=1;
  initfail(P) // Prepare fail links
  repeat
    if j==0 or $S[i] == P[j]$
      then i++, j++ // advance in text and in pattern
    else j = fail[j] // use fail link
  until j>m or i>n
  if j>m then report match at $i-m$
```
Initialization of fail links

Algorithm: KMP_Initfail

Input: Pattern P[1..m]
Output: fail[] for pattern P

\[ i=1, \ j=0, \ \text{fail}[1]=0 \]

\textbf{repeat}
  \[ \text{if } j==0 \ \text{or } P[i] == P[j] \]
  \[ \text{then } i++, \ j++, \ \text{fail}[i] = j \]
  \[ \text{else } j = \text{fail}[j] \]
\textbf{until } i>=m
Initialization of fail links

\[ i=1, \ j=0 \ , \ fail[1]= 0 \]

repeat
  \[ \text{if } j==0 \text{ or } P[i] == P[j] \]
  \[ \text{then } i++ , \ j++ , \ fail[i] = j \]
  \[ \text{else } j = fail[j] \]
until \( i>=m \)
Time complexity of KMP matching?

**Input:** Text $S[1..n]$ and pattern $P[1..m]$

**Output:** First occurrence of $P$ in $S$ (if exists)

```
i=1; j=1;
initfail(P) // Prepare fail links
repeat
    if $j==0$ or $S[i] == P[j]$
        then  $i++$, $j++$ // advance in text and in pattern
    else  $j = fail[j] // use fail link$
until  $j>m$ or $i>n$
if  $j>m$ then report match at $i-m$
```
Analysis of time complexity

• At every cycle either i and j increase by 1
• Or j decreases (j=fail[j])

• i can increase n (or m) times
• Q: How often can j decrease?
  – A: not more than nr of increases of i

• **Amortised analysis:** $O(n)$, preprocess $O(m)$
Karp-Rabin


- Compare in $O(1)$ a hash of $P$ and $S[i..i+m-1]$

- Goal: $O(n)$.

- $f( h(T[i..i+m-1]) \rightarrow h(T[i+1..i+m]) ) = O(1)$
Karp-Rabin


- Compare in $O(1)$ a hash of $P$ and $S[i..i+m-1]$

- Goal: $O(n)$.

- $f( h(T[i..i+m-1]) \rightarrow h(T[i+1..i+m]) ) = O(1)$
Hash

- “Remove” the effect of $T[i]$ and “Introduce” the effect of $T[i+m]$ – in $O(1)$

- Use base $|\Sigma|$ arithmetics and treat characters as numbers

- In case of hash match – check all $m$ positions
- Hash collisions => Worst case $O(nm)$
Let’s use numbers

• \( T = 57125677 \)
• \( P = 125 \) (and for simplicity, \( h = 125 \))

• \( H(T[1]) = 571 \)
• \( H(T[2]) = (571-5*100)*10 + 2 = 712 \)

• \( H(T[3]) = (H(T[2]) - \text{ord}(T[1])*10m)*10 + T[3+m-1] \)
hash

• $c$ – size of alphabet

• $HS_i = H(S[i..i+m-1])$

• $H(S[i+1 .. i+m]) = (HS_i - \text{ord}(S[i])* c^{m-1} ) * c + \text{ord}( S[i+m] )$

• Modulo arithmetic – to fit value in a word!
• $hash(w[0 .. m-1]) = (w[0] \times 2^{m-1} + w[1] \times 2^{m-2} + \cdots + w[m-1] \times 2^0) \mod q$
Karp-Rabin

Input: Text S[1..n] and pattern P[1..m]
Output: Occurrences of P in S
1. c=20; /* Size of the alphabet, say nr. of aminoacids */
2. q = 33554393 /* q is a prime */
3. cm = c^{m-1} \mod q
4. hp = 0 ; hs = 0
5. for i = 1 .. m do hp = ( hp*c + ord(p[i]) ) \mod q // H(P)
6. for i = 1 .. m do hs = ( hp*c + ord(s[i]) ) \mod q // H(S[1..m])
7. if hp == hs and P == S[1..m] report match at position
8. for i=2 .. n-m+1
9. hs = ( (hs - ord(s[i-1])*cm) * c + ord(s[i+m-1]) ) \mod q
10. if hp == hs and P == S[i..i+m-1]
11. report match at position i
More ways to ensure $O(n)$?
Shift-AND / Shift-OR

• Ricardo Baeza-Yates, Gaston H. Gonnet
A new approach to text searching
*Communications of the ACM* October 1992,
Volume 35 Issue 10
[ACM Digital Library: http://doi.acm.org/10.1145/135239.135243] [DOI]

• **PDF**
Bit-operations

• Maintain a set of all prefixes that have so far had a perfect match

• On the next character in text update all previous pointers to a new set

• Bit vector: for every possible character
Matching in linear time (shift-OR)

Pattern: ABCB

Text:

A

... 1 1 1 1 1 0
1 1 1 0

B

... 1 1 1 1 1 0
0 1 0 1

C

... 1 1 1 1 0 1 0
1 0 1 1

shift

| b[ T[j] ]

A

B

C

A

B

C

B

E

A

... 1 1 1 0 1 1 0
1 1 0 1

... 1 1 1 1 1 1 1
State: which prefixes match?
Shift-AND ; shift-OR
Move to next: **shift-AND**
shift 1, introduce 1, bitwise and

\[
\text{Pattern}[S[i]] \quad \& \quad 0001110001 &= 0101010001
\]
Track positions of prefix matches

Mask on char T[i]

Shift left $<<$

Bitwise AND
Vectors for every char in $\Sigma$

- $P=$aste

$$
\begin{align*}
 \text{a s t e b c d .. z} \\
1000000 ... \\
0100000 ... \\
0010000 ... \\
0001000 ... \\
0000100 ... \\
0000010 ... \\
\end{align*}
$$
• $T = \text{lasteaed}$

lasteaed

0 1
0 0
0 0
0 0
0 0
• \( T = \text{lasteaed} \)

\[
\begin{align*}
\text{lasteaed} \\
0 & 1 & 0 \\
0 & 0 & 1 \\
0 & 0 & 0 \\
0 & 0 & 0
\end{align*}
\]
• $T = \text{lasteaed}$

\[
\begin{array}{cccccc}
  & 0 & 1 & 0 & 0 & 0 & 1 \\
 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & 0 \\
\end{array}
\]
• $T = \text{lasteaeed}$

```
l a s t e a e d
0 1 0 0 0 1
0 0 1 0 0 0
0 0 0 1 0 0
0 0 0 0 1 0
0 0 0 0 0 1
```
Description

The Shift Or algorithm uses bitwise techniques. Let $R$ be a bit array of size $m$. Vector $R_j$ is the value of the array $R$ after text character $y[j]$ has been processed (see figure 5.1). It contains informations about all matches of prefixes of $x$ that end at position $j$ in the text for $0 < i \leq m-1$:

$$R_j[i] = \begin{cases} 
0 & \text{if } x[0..i] = y[j-i..j], \\
1 & \text{otherwise.}
\end{cases}$$

![Figure 5.1: Meaning of vector $R_j$.](http://www-igm.univ-mlv.fr/~lecroq/string/node6.html)

The vector $R_{j+1}$ can be computed after $R_j$ as follows. For each $R_j[i] = 0$:

$$R_{j+1}[i + 1] = \begin{cases} 
0 & \text{if } x[i + 1] = y[j + 1], \\
1 & \text{otherwise,}
\end{cases}$$

and

$$R_{j+1}[0] = \begin{cases} 
0 & \text{if } x[0] = y[j + 1], \\
1 & \text{otherwise.}
\end{cases}$$

If $R_{j+1}[m-1] = 0$ then a complete match can be reported.
The C code

```c
int preSo(char *x, int m, unsigned int S[]) {
    unsigned int j, lim;
    int i;
    for (i = 0; i < ASIZE; ++i)
        S[i] = 0;
    for (lim = i = 0, j = 1; i < m; ++i, j <<= 1) {
        S[x[i]] &= ~j;
        lim |= j;
    }
    lim = ~(lim>>1);
    return(lim);
}

void SO(char *x, int m, char *y, int n) {
    unsigned int lim, state;
    unsigned int S[ASIZE];
    int j;
    if (m > WORD)
        error("SO: Use pattern size <= word size");

    /* Preprocessing */
    lim = preSo(x, m, S);

    /* Searching */
    for (state = 0, j = 0; j < n; ++j) {
        state = (state<<1) | S[y[j]];
        if (state < lim)
            OUTPUT(j - m + 1);
    }
}
```
The example

Searching phase

<table>
<thead>
<tr>
<th></th>
<th>[A]</th>
<th>1</th>
<th>1</th>
<th>0</th>
<th>1</th>
<th>0</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>G</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>C</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>A</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>G</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>A</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>G</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>A</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>G</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

As $R_{12}[7]=0$ it means that an occurrence of $x$ has been found at position $12-8+1=5$. 
## Summary

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Worst case</th>
<th>Ave. Case</th>
<th>Preprocess</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brute force</td>
<td>O(mn)</td>
<td>O(n * (1+1/</td>
<td>Σ</td>
</tr>
<tr>
<td>Knuth-Morris-Pratt</td>
<td>O(n)</td>
<td>O(n)</td>
<td>O(m)</td>
</tr>
<tr>
<td>Rabin-Karp</td>
<td>O(mn)</td>
<td>O(n)</td>
<td>O(m)</td>
</tr>
<tr>
<td>Boyer-Moore</td>
<td></td>
<td>O(n/m) ?</td>
<td></td>
</tr>
<tr>
<td>BM Horspool</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Factor search</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Shift-OR</td>
<td>O(n)</td>
<td>O(n)</td>
<td>O( m</td>
</tr>
</tbody>
</table>
Fig. 2.22. Map of experimental efficiency for different string matching algorithms.
• R. Boyer, S. Moore: A fast string searching algorithm. *CACM* 20 (1977), 762-772 [PDF]
Find occurrences in text

- Have we missed anything?
Find occurrences in text

• What have we learned if we test for a potential match from the end?
Our search algorithm may be specified as follows:

\[ \text{stringlen} \leftarrow \text{length of string}. \]
\[ i \leftarrow \text{patlen}. \]

**top**: if \( i > \text{stringlen} \) then return false.
\[ j \leftarrow \text{patlen}. \]

**loop**: if \( j = 0 \) then return \( j + 1 \).
\[ \text{if string}(i) = \text{pat}(j) \]
\[ \quad \text{then} \]
\[ \quad j \leftarrow j - 1. \]
\[ \quad i \leftarrow i - 1. \]
\[ \quad \text{goto loop.} \]
\[ \text{close;} \]
\[ i \leftarrow i + \max(\text{delta}_1 (\text{string}(i)), \text{delta}_2 (j)). \]
\[ \text{goto top.} \]

If the above algorithm returns false, then \( \text{pat} \) does not occur in \( \text{string} \). If the algorithm returns a number, then it is the position of the left end of the first occurrence of \( \text{pat} \) in \( \text{string} \).
Find occurrences in text
Bad character heuristics
maximal shift on $S[i]$

$$\delta_1(S[i]) = \begin{cases} |m| & \text{if pattern does not contain } S[i] \\ \text{patlen-j} & \text{max } j \text{ so that } P[j] == S[i] \end{cases}$$

First x in pattern (from end)
void bmInitocc() {
    char a; int j;
    for(a=0; a<alphabetsize; a++)
        occ[a]=-1;
    for (j=0; j<m; j++) {
        a=p[j];
        occ[a]=j; }
}
Good suffix heuristics

\[
\delta_2(S[i]) - \text{minimal shift so that matched region is fully covered or that the suffix of match is also a prefix of P}
\]
Boyer-Moore algorithm

Input: Text $S[1..n]$ and pattern $P[1..m]$  
Output: Occurrences of $P$ in $S$

preprocess_BM() // delta1 and delta2  
i=m  
while $i \leq n$  
  for( $j=m$; $j>0$ and $P[j]==S[i-m+j]$; $j--$ ) ;  
  if $j==0$ report match at position $i-m+1$  
  $i = i + \max( \text{delta1}[ S[i] ], \text{delta2}[ j ] )$
• http://www.iti.fh-flensburg.de/lang/algorithmen/pattern/bmen.htm


• Animation: http://www-igm.univ-mlv.fr/~lecroq/string/
Simplifications of BM

• There are many variants of Boyer-Moore, and many scientific papers.
• On average the time complexity is sublinear.
• Algorithm speed can be improved and yet simplify the code.
• It is useful to use the last character heuristics (Horspool (1980), Baeza-Yates(1989), Hume and Sunday(1991)).
Algorithm BMH (Boyer-Moore-Horspool)

• RN Horspool - Practical Fast Searching in Strings
  Software - Practice and Experience, 10(6):501-506 1980

Input: Text S[1..n] and pattern P[1..m]
Output: occurrences of P in S
1. for a in Σ do delta[a] = m
2. for j=1..m-1 do delta[P[j]] = m-j
3. i=m
4. while i <= n
5.    if S[i] == P[m]
6.       j = m-1
7.    while (j>0 and P[j]==S[i-m+j]) j = j-1 ;
8.    if j==0 report match at i-m+1
9.    i = i + delta[S[i]]
String Matching: Horspool algorithm

• How the comparison is made?

Text:

Pattern:

From right to left: suffix search

• Which is the next position of the window?

Text:

Pattern:

It depends of where appears the last letter of the text, say it ‘a’, in the pattern:

Then it is necessary a preprocess that determines the length of the shift.
Algorithm Boyer-Moore-Horspool-Hume-Hume-Sunday (BMHHS)

- Use delta in a tight loop
- If match (delta==0) then check and apply original delta d

Input: Text S[1..n] and pattern P[1..m]
Output: occurrences of P in S

1. for a in Σ do delta[a] = m
2. for j=1..m-1 do delta[P[j]] = m-j
3. d = delta[ P[ m ] ];       // memorize d on P[m]
4. delta[ P[ m ] ] = 0;      // ensure delta on match of last char is 0
5. for ( i=m ; i<= n ; i = i+d )
6.   repeat                     // skip loop
7.       t=delta[ S[i] ] ; i = i + t
8.   until t==0
9.   for( j=m-1 ; j> 0 and P[j]==S[i-m+j] ; j = j-1 ) ;
10.  if j==0 report match at i-m+1

BMHHS requires that the text is padded by P: S[n+1]..S[n+m] = P
(in order for the algorithm to finish correctly – at least one occurrence!).
• **Daniel M. Sunday:** A very fast substring search algorithm [PDF]
  *Communications of the ACM August 1990, Volume 33 Issue 8*

• **Loop unrolling:**
  • Avoid too many loops (each loop requires tests) by just repeating code within the loop.
  • Line 7 in previous algorithm can be replaced by:

    7. i += delta[ S[i ]];
    i += delta[ S[i ]];
    i += (t = delta[ S[i ]]) ;
2.1.5 Popularity

Although providing a high performance, the degree to which the Boyer-Moore approach has been put into practice may have been curbed to a certain extent by conceptual difficulties in the preprocessing, particularly with the match heuristic. Horspool (1980) observed that "many programmers may not believe that the Boyer and Moore algorithm (if they have heard of it) is a truly practical approach." Sedgewick (1983) also noted that "both the Knuth-Morris-Pratt and the Boyer-Moore algorithms require some complicated preprocessing on the pattern that is difficult to understand and has limited the extent to which they are used." In fact, Morris discovered that after a few months his initial search implementation in a text editor, being too difficult to understand by other systems programmers, had been ruined by various gratuitous 'fixes' (Knuth, Morris and Pratt, 1977).

Hume and Sunday (1991) add that "partially because the best algorithms presented in the literature are difficult to understand and to implement, knowledge of fast and practical algorithms is not commonplace." Arguing for a more consistent approach to algorithm development, and against "widespread chaotic algorithm presentation," Woude (1989) asserts that the Knuth-Morris-Pratt preprocessing should be no more difficult to understand than the actual search procedure itself. There have also been some attempts of late to popularise the Boyer-Moore approach, particularly in its simplified form (e.g. Menico, 1989). As a step towards elucidating this field, Hume and Sunday (1991) have presented a means of classifying the various string-matching algorithms based on the Knuth-Morris-Pratt and Boyer-Moore approaches. A taxonomy of string-matching algorithms, including these two aforementioned approaches, has also been put forward by Watson and Zwaan (1992).
Forward-Fast-Search: Another Fast Variant of the Boyer-Moore String Matching Algorithm

• The Prague Stringology Conference '03
• Domenico Cantone and Simone Faro
• Abstract: We present a variation of the Fast-Search string matching algorithm, a recent member of the large family of Boyer-Moore-like algorithms, and we compare it with some of the most effective string matching algorithms, such as Horspool, Quick Search, Tuned Boyer-Moore, Reverse Factor, Berry-Ravindran, and Fast-Search itself. All algorithms are compared in terms of run-time efficiency, number of text character inspections, and number of character comparisons. It turns out that our new proposed variant, though not linear, achieves very good results especially in the case of very short patterns or small alphabets.
• http://cs.felk.cvut.cz/psc/event/2003/p2.html
• PS.gz (local copy)
Factor based approach

• Optimal average-case algorithms
  – Assuming independent characters, same probability

• Factor – a substring of a pattern
  – Any substring
  – (how many?)
Factor based approach

simple. It is shown in Figure 2.13. Suppose that we have read backward a factor $u$ of the pattern, and that we failed on the next letter $\sigma$. This means that the string $\sigma u$ is no longer a factor of $p$, so no occurrence of $p$ can contain $\sigma u$, and we can safely shift the window to after $\sigma$.

Fig. 2.13. Basic idea for shifting the window with the factor search approach. If we failed to recognize a factor of the pattern on $\sigma$, then $\sigma u$ is not a factor of the pattern and the window can be safely shifted after $\sigma$. 
Factor searches

Do not compare characters, but find the longest match to any subregion of the pattern.
Examples

• Backward DAWG Matching (BDM)
  – Crochemore et al 1994

• Backward Nondeterministic DAWG Matching (BNDM)
  – Navarro, Raffinot 2000

• Backward Oracle Matching (BOM)
  – Allauzen, Crochemore, Raffinot 2001
Backward DAWG Matching BDM

Suffix automaton recognises all factors (and suffixes) in $O(n)$

Fig. 2.14. Basic search of the BDM algorithm with the suffix automaton. The variable \textit{last} stores the beginning position of the longest suffix of the part read that is also a prefix of the pattern.

Do not compare characters, but find the longest match to any subregion of the pattern.
BNDM – simulate using bitparallelism

Bits – show where the factors have occurred so far

String matching

with 0 and 1 as with Shift-And. The number 1, representing an active state at position \( j \) of \( p \), means that the factor \( p_{j \ldots j+|u|-1} \) is equal to \( u \). Figure 2.15 shows this relationship. If the pattern is of size less than \( w \), then the set fits in a computer word \( D = d_m \ldots d_1 \).

![Diagram showing bit-parallel factor search.]

Fig. 2.15. Bit-parallel factor search. The table \( D \) keeps a list of the positions in \( p \) where the factor \( u \) begins.
BNDM matches an NDA

NDA on the suffixes of ‘announce’

Fig. 2.17. Nondeterministic automaton recognizing all factors of the reverse string of “announce”.
Deterministic version of the same Backward Factor Oracle

Fig. 2.20. Factor oracle for the reverse string of “announce”.
2.6 Other algorithms and references

Fig. 2.22. Map of experimental efficiency for different string matching algorithms.

BNDM – Backward Non-Deterministic DAWG Matching
BOM - Backward Oracle matching
String Matching of one pattern

CTACTACTACGTCTATACTGATCGTAGC
TACTACGGGTATGACTAA

1. Prefix search
2. Suffix search
3. Factor search
Multiple patterns

S

\{P\}
Why?

• Multiple patterns
  • Highlight multiple different search words on the page
  • Virus detection – filter for virus signatures
  • Spam filters
  • Scanner in compiler needs to search for multiple keywords
  • Filter out stop words or disallowed words
  • Intrusion detection software
  • Next-generation sequencing produces huge amounts (many millions) of short reads (20-100 bp) that need to be mapped to genome!
• ...
Algorithms

• Aho-Corasick (search for multiple words)
  – Generalization of Knuth-Morris-Pratt
• Commentz-Walter
  – Generalization of Boyer-Moore & AC
• Wu and Manber
  – improvement over C-W

• Additional methods, tricks and techniques
Aho-Corasick (AC)


• ACM:DOI PDF

• ABSTRACT This paper describes a simple, efficient algorithm to locate all occurrences of any of a finite number of keywords in a string of text. The algorithm consists of constructing a finite state pattern matching machine from the keywords and then using the pattern matching machine to process the text string in a single pass. Construction of the pattern matching machine takes time proportional to the sum of the lengths of the keywords. The number of state transitions made by the pattern matching machine in processing the text string is independent of the number of keywords. The algorithm has been used to improve the speed of a library bibliographic search program by a factor of 5 to 10.

• Generalization of KMP for many patterns
• Text $S$ like before.
• Set of patterns $\mathbf{P} = \{ P_1, \ldots, P_k \}$
• Total length $|\mathbf{P}| = m = \sum_{i=1}^{k} m_i$
• Problem: find all occurrences of any of the $P_i \in \mathbf{P}$ from $S$
Idea

1. Create an automaton from all patterns

2. Match the automaton

• Use the PATRICIA trie for creating the main structure of the automaton
PATRICIA trie

- **Abstract** PATRICIA is an algorithm which provides a flexible means of storing, indexing, and retrieving information in a large file, which is economical of index space and of reindexing time. It does not require rearrangement of text or index as new material is added. It requires a minimum restriction of format of text and of keys; it is extremely flexible in the variety of keys it will respond to. It retrieves information in response to keys furnished by the user with a quantity of computation which has a bound which depends linearly on the length of keys and the number of their proper occurrences and is otherwise independent of the size of the library. It has been implemented in several variations as FORTRAN programs for the CDC-3600, utilizing disk file storage of text. It has been applied to several large information-retrieval problems and will be applied to others.

- [ACM:DOI PDF](#)
Trie

From Wikipedia, the free encyclopedia

In computer science, a trie, or prefix tree, is an ordered tree data structure that is used to store an associative array where the keys are usually strings. Unlike a binary search tree, no node in the tree stores the key associated with that node; instead, its position in the tree shows what key it is associated with. All the descendants of any one node have a common prefix of the string associated with that node, and the root is associated with the empty string. Values are normally not associated with every node, only with leaves and some inner nodes that happen to correspond to keys of interest.

The term trie comes from "retrieval." Following the etymology, the inventor, Edward Fredkin, pronounces it [tri] ("tree") [1]. However, it is usually pronounced [trai] ("try").[2]
• **Word trie** - a good data structure to represent a set of words (e.g. a dictionary).
• **trie** (data structure)

• **Definition:** A tree for storing strings in which there is one node for every common prefix. The strings are stored in extra leaf nodes.

• See also digital tree, digital search tree, directed acyclic word graph, compact DAWG, Patricia tree, suffix tree.

• **Note:** The name comes from retrieval and is pronounced, "tree."
• To test for a word p, only $O(|p|)$ time is used no matter how many words are in the dictionary...
Trie for P=\{he, she, his, hers\}
Trie for \( P=\{\text{he, she, his, hers}\} \)
How to search for words like he, sheila, hi. Do these occur in the trie?

\[ \text{trie diagram} \]
Aho-Corasick

1. Create an automaton $M_P$ for a set of strings $P$.
2. Finite state machine: read a character from text, and change the state of the automaton based on the state transitions...
3. Main links: $\text{goto}[j,c]$ - read a character $c$ from text and go from a state $j$ to state $\text{goto}[j,c]$.
4. If there are no $\text{goto}[j,c]$ links on character $c$ from state $j$, use $\text{fail}[j]$.
5. Report the output. Report all words that have been found in state $j$. 
AC Automaton (vs KMP)

goto[1,i] = 6. ;
fail[7] = 3,
fail[8] = 0 ,

Output table
state output[j]
2 he
5 she, he
7 his
9 hers

NOT { h, s }
AC - matching

**Input**: Text $S[1..n]$ and an AC automaton $M$ for pattern set $P$

**Output**: Occurrences of patterns from $P$ in $S$ (last position)

1. $state = 0$

2. for $i = 1..n$ do

3. \hspace{1cm} while $(\text{goto}[state, S[i]] == \emptyset) \text{ and } (\text{fail}[state] != state)$ do

4. \hspace{1cm} $state = \text{fail}[state]$

5. \hspace{1cm} $state = \text{goto}[state, S[i]]$

6. \hspace{1cm} if ( output[state] not empty )

7. \hspace{1cm} then report matches output[state] at position $i$
Algorithm Aho-Corasick preprocessing I (TRIE)

Input: \( P = \{ P_1, \ldots, P_k \} \)
Output: goto[] and partial output[]
Assume: output(s) is empty when a state s is created;
goto[s,a] is not defined.

```
procedure enter(a_1, \ldots, a_m) /* P_i = a_1, \ldots, a_m */
begin
1. s=0; j=1;
2. while goto[s,a_j] ≠ ∅ do // follow existing path
   3. \( s = \text{goto}[s,a_j] \);
   4. \( j = j+1 \);
5. for p=j to m do // add new path (states)
   6. news = news+1 ;
   7. goto[s,a_p] = news;
   8. s = news;
9. output[s] = a_1, \ldots, a_m
end
```

begin
10. news = 0
11. for i=1 to k do enter( P_i )
12. for a ∈ Σ do
13. if goto[0,a] = ∅ then goto[0,a] = 0 ;
end
Preprocessing II for AC (FAIL)

queue = ∅
for a ∈ Σ do
    if goto[0,a] ≠ 0 then
        enqueue( queue, goto[0,a] )
        fail[ goto[0,a] ] = 0

while queue ≠ ∅
    r = take( queue )
    for a ∈ Σ do
        if goto[r,a] ≠ ∅ then s = goto[ r, a ]
        enqueue( queue, s ) // breadth first search
        state = fail[r]
        while goto[state,a] = ∅ do state = fail[state]
        fail[s] = goto[state,a]
        output[s] = output[s] + output[ fail[s] ]
Correctness

• Let string \( t \) "point" from initial state to state \( j \).

• Must show that \( \text{fail}[j] \) points to longest suffix that is also a prefix of some word in \( P \).

• Look at the article...
AC matching time complexity

- **Theorem** For matching the $M_p$ on text $S$, $|S|=n$, less than $2n$ transitions within $M$ are made.
- **Proof** Compare to KMP.
- There is at most $n$ goto steps.
- Cannot be more than $n$ Fail-steps.
- In total -- there can be less than $2n$ transitions in $M$. 
Individual node (goto)

- Full table
- List
- Binary search tree(?)
- Some other index?
AC thoughts

• Scales for many strings simultaneously.
• For very many patterns – search time (of grep) improves(??)
  – See Wu-Manber article
• When k grows, then more fail[] transitions are made (why?)
• But always less than n.
• If all goto[j,a] are indexed in an array, then the size is |M_p|*|Σ|, and the running time of AC is O(n).
• When k and c are big, one can use lists or trees for storing transition functions.
• Then, $O(n \log(\min(k,c)))$.
Advanced AC

• Precalculate the next state transition correctly for every possible character in alphabet

• Can be good for short patterns
Problems of AC?

- Need to rebuild on adding / removing patterns
- Details of branching on each node(?)
Commentz-Walter

• Generalization of Boyer-Moore for multiple sequence search

• Beate Commentz-Walter
  A String Matching Algorithm Fast on the Average
  Proceedings of the 6th Colloquium, on Automata, Languages and Programming. Lecture Notes In

Aho and Corasick [AC75] presented a linear-time algorithm for this problem, based on an automata approach. This algorithm serves as the basis for the UNIX tool fgrep. A linear-time algorithm is optimal in the worst case, but as the regular string-searching algorithm by Boyer and Moore [BM77] demonstrated, it is possible to actually skip a large portion of the text while searching, leading to faster than linear algorithms in the average case.
Commentz-Walter [CW79]

• Commentz-Walter [CW79] presented an algorithm for the multi-pattern matching problem that combines the Boyer-Moore technique with the Aho-Corasick algorithm. The Commentz-Walter algorithm is substantially faster than the Aho-Corasick algorithm in practice. Hume [Hu91] designed a tool called gre based on this algorithm, and version 2.0 of fgrep by the GNU project [Ha93] is using it.

• Baeza-Yates [Ba89] also gave an algorithm that combines the Boyer-Moore-Horspool algorithm [Ho80] (which is a slight variation of the classical Boyer-Moore algorithm) with the Aho-Corasick algorithm.
Idea of C-W

• Build a **backward** trie of all keywords

• Match from the end until mismatch...

• Determine the shift based on the combination of heuristics
Horspool for many patterns

Search for ATGTATG, TATG, ATAAT, ATGTG

1. Build the trie of the inverted patterns

2. $l_{\text{min}}=4$

3. Table of shifts

4. Start the search

Slides courtesy of: Xavier Messeguer Peypoch (http://www.lsi.upc.es/~alggen)
Horspool for many patterns

Search for ATGTATG, TATG, ATAAT, ATGTG

The text ACATGCTATGTGACA...

Slides courtesy of: Xavier Meseguer Peypoch (http://www.lsi.upc.es/~alggen)
Horspool for many patterns

Search for ATGTATG, TATG, ATAAT, ATGTG

The text ACATGCTATGTGACA...

Slides courtesy of: Xavier Meseguer Peypoch (http://www.lsi.upc.es/~alggen)
Horspool for many patterns

Search for ATGTATG, TATG, ATAAT, ATGTG

The text ACATGCTATGTGACA...

Slides courtesy of: Xavier Messeguer Peypoch (http://www.lsi.upc.es/~alggen)
Horspool for many patterns

Search for ATGTATG, TATG,ATAAT, ATGTG

The text ACATGCTATGTGACA...
Horspool for many patterns

Search for ATGTATG, TATG, ATAAT, ATGTG

The text ACATGCTATGTGACA...

Slides courtesy of: Xavier Messeguer Peypoch (http://wwwlsi.upc.es/~alggen)
Horspool for many patterns

Search for ATGTATG, TATG, ATAAT, ATGTG

The text ACATGCTATGTGACA...

Slides courtesy of: Xavier Messeguer Peypoch (http://www.lsi.upc.es/~alggen)
Horspool for many patterns

Search for ATGTATG, TATG, ATAAT, ATGTG

The text ACATGCTATGTGACA...

Short Shifts!

Slides courtesy of: Xavier Messeguer Peypoch (http://www.lsi.upc.es/~alggen)
What are the possible limitations for C-W?

• Many patterns, small alphabet – minimal skips

• What can be done differently?
Wu-Manber

- Citeseer: http://citeseer.ist.psu.edu/wu94fast.html  [Postscript]
- We present a different approach that also uses the ideas of Boyer and Moore. Our algorithm is quite simple, and the main engine of it is given later in the paper. An earlier version of this algorithm was part of the second version of agrep [WM92a, WM92b], although the algorithm has not been discussed in [WM92b] and only briefly in [WM92a]. The current version is used in glimpse [MW94]. **The design of the algorithm concentrates on typical searches** rather than on worst-case behavior. This allows us to **make some engineering decisions** that we believe are crucial to making the algorithm significantly faster than other algorithms in practice.
Key idea

• Main problem with Boyer-Moore and many patterns is that, the more there are patterns, the shorter become the possible shifts...

• Wu and Manber: check several characters simultaneously, i.e. increase the alphabet.
Instead of looking at characters from the text one by one, we consider them in blocks of size $B$.

$log_c 2M$; in practice, we use either $B = 2$ or $B = 3$.

The SHIFT table plays the same role as in the regular Boyer-Moore algorithm, except that it determines the shift based on the last $B$ characters rather than just one character.
Horspool to Wu-Manber

How do we can increase the length of the shifts?

With a table shift of l-mers with the patterns ATGTATG, TATG, ATAAT, ATGTG

<table>
<thead>
<tr>
<th>1 símbol</th>
<th>2 símbols</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1</td>
</tr>
<tr>
<td>C</td>
<td>4 (lmin)</td>
</tr>
<tr>
<td>G</td>
<td>2</td>
</tr>
<tr>
<td>T</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>2 símbols</th>
</tr>
</thead>
<tbody>
<tr>
<td>AA</td>
</tr>
<tr>
<td>AC</td>
</tr>
<tr>
<td>AG</td>
</tr>
<tr>
<td>AT</td>
</tr>
<tr>
<td>CA</td>
</tr>
<tr>
<td>CC</td>
</tr>
<tr>
<td>CG</td>
</tr>
<tr>
<td>...</td>
</tr>
</tbody>
</table>

Slides courtesy of: Xavier Messeguer Peypoch (http://www.lsi.upc.es/~alggen)
Wu-Manber algorithm

Search for ATGTATG, TATG, ATAAT, ATGTG

into the text: ACATGCTATGTGACATAATA

Experimental length: $\log_{|\Sigma|} 2^{l_{\text{min}}*r}$

Slides courtesy of: Xavier Messeguer Peypoch (http://wwwlsi.upc.es/~alggen)
Backward Oracle

• Set Backwards oracle SBDM, SBOM

• Pages 68-72
3.4.3 Set Backward Oracle Matching algorithm

The Set Backward Oracle Matching algorithm (SBOM) [AR99] uses a factor oracle of the set of strings. The factor oracle of \( P \) recognizes at least all the factors of the strings in \( P \). The search algorithm is similar to SDBM. We slide a window of size \( l_{\text{min}} \) along the text, reading backward a suffix of the window in the factor oracle. If we fail on a letter \( \sigma \), we can safely shift the window past \( \sigma \). If not, we reach the beginning of the window and verify a subset of \( P \) against the text.

3.4.3.1 Factor oracle of a set of strings

The factor oracle construction on a set of strings resembles the Aho-Corasick automaton construction. The only difference appears when going down the supply path looking for an outgoing transition labeled by \( \sigma \). In the Aho-Corasick automaton construction, if this transition does not exist, we just jump to the next state on the supply path (Section 3.2.2). In the factor oracle construction, we create in addition a transition labeled by \( \sigma \) from each state on the supply path to the state where the original transition leads.
String matching of many patterns

Wu-Manber

SBOM

<table>
<thead>
<tr>
<th>Σ</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
</tr>
<tr>
<td>6</td>
</tr>
<tr>
<td>4</td>
</tr>
<tr>
<td>2</td>
</tr>
</tbody>
</table>

Lmin

5 10 15 20 25 30 35 40 45

Slides courtesy of: Xavier Messeguer Peypoch (http://www.lsi.upc.es/~alggen)
String matching of many patterns

Wu-Manber

<table>
<thead>
<tr>
<th>Σ</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
</tr>
<tr>
<td>4</td>
</tr>
<tr>
<td>2</td>
</tr>
</tbody>
</table>

SBOM

Lmin

5              10              15              20            25             30            35           40            45

(5 patterns)

(10 patterns)

(100 patterns)

(1000 patterns)

Slides courtesy of: Xavier Messeguer Peypoch (http://www.lsi.upc.es/~alggen)
string matching algorithms, showing for all of them the zone in which they are most efficient in practice. The text of 10 megabytes is randomly built, as are the patterns. The experiments were performed on a \( w = 32 \) bits Ultra Sparc 1 running SunOS 5.6. The sets contain 5, 10, 100, and 1000 strings of the same length, varying from 5 to 100 in steps of 5. We tested all the algorithms presented. The Wu-Manber algorithm used in these experiments is the implementation found in Agrep. Its performance may vary, depending on the hash functions and the sizes of the tables used.

Fig. 3.24. Map of the most efficient algorithms when searching for 5 strings.

The maps are shown in Figures 3.24 to 3.27. The most efficient algorithms are just Wu-Manber, the advanced Aho-Corasick, and SBOM. As the set grows in size, SBOM becomes more and more attractive. The advanced Aho-Corasick also improves in comparison with the others for short strings, since it reads the text only once.
10 strings

Fig. 3.25. Map of the most efficient algorithms when searching for 10 strings.
Fig. 3.26. Map of the most efficient algorithms when searching for 100 strings.
1000 strings

Fig. 3.27. Map of the most efficient algorithms when searching for 1000 strings.

On matching a set of strings on unbounded alphabets The problem of matching a set of strings of the same length $m$ on an unbounded alphabet has been investigated [Bre95]. The resulting algorithm runs in $O((\log(|P|)/m + 1) \times n)$ comparisons after an $O(|P| \times m \times \log |A|)$ preprocessing time, where $A$ is the alphabet on which the set $P$ is built.
Factor Oracle
Factor Oracle: safe shift
Factor Oracle:

Shift to match prefix of P2?
3.4 Factor based approach

Example using DNA. We search for the set of strings \( P = \{\text{ATATATA, TATAT, ACAGAT}\} \) in the text AGATACGATATATAC. The factor oracle for the reverse set of \( P_{\text{min}} = \{\text{ATATA, TATAT, ACAGA}\} \) is shown in Figure 3.23.

Fig. 3.23. Factor oracle for the reverse set of \( P_{\text{min}} = \{\text{ATATA, TATAT, ACAGA}\} \). Double-circled states are terminal.
Construction of factor Oracle

The resulting factor automaton recognizes at least all factors of $P$ [AR99]. The construction algorithm is worst-case time $O(|P|)$. Its pseudo-code is given in Figure 3.20.

```
Build_Oracle_Multiple(P = \{p^1, p^2, \ldots, p^n\})
1. OR.trie ← Trie(P)
   δOR is its transition function
2. Mark the states that correspond to an entire string $p^i$ as terminal
3. I ← root of OR.trie
4. SOR(I) ← θ
5. For Current in transversal order Do
   Parent ← parent in OR.trie of Current
   σ ← label of the transition from Parent to Current
   Down ← SOR(Parent)
   While Down ≠ θ AND δOR(Down, σ) = θ Do
   δOR(Down, σ) ← Current
   Down ← SOR(Down)
   End of while
   If Down ≠ θ Then
   SOR(Current) ← δOR(Down, σ)
   Else SOR(Current) ← I
   End of if
7. End of for
```

Fig. 3.20. Construction of the factor oracle for a set $P = \{p^1, p^2, \ldots, p^n\}$. The state $Current$ goes through the trie $OR.trie$ built on $P$ in transversal order. The state $Down$ goes down the supply links from the parent of $Current$ looking for an outgoing transition labeled with the same character as between $Current$ and its parent, creating it if it does not exist.
Factor oracle


• http://portal.acm.org/citation.cfm?id=647009.712672&coll=GUIDE&dl=GUIDE&CFID=31549541&CFTOKEN=61811641#

• http://www-igm.univ-mlv.fr/~allauzen/work/sofsem.ps
2 Factor oracle

2.1 Construction algorithm

Build\_Oracle\((p = p_1p_2 \ldots p_m)\)
1. For \(i\) from 0 to \(m\)
2. Create a new state \(i\)
3. For \(i\) from 0 to \(m - 1\)
4. Build a new transition from \(i\) to \(i + 1\) by \(p_{i+1}\)
5. For \(i\) from 0 to \(m - 1\)
6. Let \(u\) be a minimal length word in state \(i\)
7. For all \(\sigma \in \Sigma, \sigma \neq p_{i+1}\)
8. If \(u\sigma \in \text{Fact}(p_{i-|u|+1} \ldots p_m)\)
9. Build a new transition from \(i\) to \(i + \text{poccur}(u\sigma; p_{i-|u|+1} \ldots p_m)\) by \(\sigma\)

Figure 1. High-level construction algorithm of the Oracle
**Definition 1** The factor oracle of a word $p = p_1p_2 \ldots p_m$ is the automaton build by the algorithm Build_Oracle (figure 1) on the word $p$, where all the states are terminal. It is denoted by $\text{Oracle}(p)$.

The factor oracle of the word $p = abbaab$ is given as an example figure 2. On this example, it can be noticed that the word $aba$ is recognized whereas it is not a factor of $p$.

![Graph](image)

**Figure2.** Factor oracle of $abbaab$. The word $aba$ is recognizes whereas it is not a factor.

Note: all the transitions that reach state $i$ of $\text{Oracle}(p)$ are labeled by $p_i$.

**Lemma 1** Let $u \in \Sigma^*$ be a minimal length word among the words recognized in state $i$ of $\text{Oracle}(p)$. Then, $u \in \text{Fact}(p)$ and $i = \text{poccur}(u, p)$. 
So far

- Generalised KMP -> AhoCorasick
- Generalised Horspool -> CommentzWalter, WuManber
- BDM, BOM
  -> Set Backward Oracle Matching...

- Other generalisations?
Multiple Shift-AND

- $P = \{P_1, P_2, P_3, P_4\}$. Generalize Shift-AND

- Bits =

- Start =

- Match =