String Matching

A Lecture in CE Freshman Seminar Series:
Ten Puzzling Problems in Computer Engineering

About This Presentation

This presentation belongs to the lecture series entitled "Ten Puzzling Problems in Computer Engineering," devised for a ten-week, one-unit, freshman seminar course by Behrooz Parhami, Professor of Computer Engineering at University of California, Santa Barbara. The material can be used freely in teaching and other educational settings. Unauthorized uses, including any use for financial gain, are prohibited.

Edition	Released	Revised	Revised	Revised	Revised
First	May 2007	May 2008	May 2009	May 2010	May 2011
		May 2012	May 2015	May 2016	May 2020

Word Search Puzzles

Type 1, With Word List Supplied

S	E	A	L	P	V	C	A	T	P	A
W	A	S	A	B	A	C	G	U	G	L
I	D	S	P	W	L	F	L	I	V	I
T	A	E	A	C	V	L	T	M	O	N
C	S	M	W	T	E	A	L	E	A	T
H	X	B	W	Y	T	J	P	U	T	S
P	T	L	Q	O	H	C	T	U	L	C
X	R	Y	R	R	E	D	I	L	G	R
Z	S	L	O	R	T	N	O	C	T	E
G	N	I	L	P	U	O	C	Y	K	E
C	O	N	N	E	C	T	O	R	S	N

AGITATOR
ASSEMBLY
CLUTCH
CONNECTORS
CONTROL
COUPLING
GLIDE
LINT SCREEN
PULLEY
SEAL
SWITCH
VALVE

The puzzle below is a little harder than the normal word search: one of the 36 first/last names has been left out (which one?)

Y	O	U	R	M	O	J	D	Z	E	R	N	A	S	O
S	U	S	A	N	A	R	A	B	R	A	B	H	T	R
R	T	R	V	B	E	H	N	N	Y	W	D	L	O	G
U	K	C	A	L	R	I	A	L	B	K	R	N	L	J
F	S	L	H	I	G	Y	L	M	E	P	A	U	L	A
M	P	A	S	A	Y	O	M	V	M	A	H	G	I	V
A	M	C	V	N	R	U	I	A	Y	P	C	U	L	S
H	K	A	O	A	B	N	E	T	S	R	I	K	A	Z
G	L	T	C	F	G	G	O	S	R	T	R	V	P	E
N	U	E	A	U	V	E	O	E	S	A	E	A	A	R
I	F	L	E	R	B	R	Z	B	R	U	C	E	L	N
B	P	L	B	E	K	I	M	M	E	L	L	I	L	E
N	O	S	L	E	N	O	J	E	R	B	A	K	E	R

AMY STEEL
KEVIN BLAIR
RON PALILLO
BARBARA BINGHAM
KIRSTEN BAKER
SHAVAR ROSS
BRUCE MAHLER
LARRY ZERNER
STU CHARNO
CAROL LACATELL
MARK NELSON
SUSAN BLU
DANA KIMMELL
PAUL KRATKA
TONY GOLDWYN
JOHN FUREY
RICHARD YOUNG
TRACIE SAVAGE

“Ten Puzzling Problems in Computer Engineering” Word Search

WORD LIST:

BINARY SEARCH
BYZANTINE GENERALS
CRYPTOGRAPHY
EASY HARD IMPOSSIBLE
MALFUNCTION DIAGNOSIS
PLACEMENT AND ROUTING
SATISFIABILITY
SORTING NETWORKS
STRING MATCHING
TASK SCHEDULING

V	K	T	A	S	K	S	C	H	E	D	U	L	I	N	G	J	Y	Y	M
L	Y	O	W	F	H	L	X	E	Z	P	X	F	L	O	N	T	D	A	G
L	E	A	S	Y	H	A	R	D	I	M	P	O	S	S	I	B	L	E	Y
O	O	F	B	H	Q	R	J	W	H	S	B	F	N	L	T	F	T	H	B
X	M	G	Y	L	D	E	R	S	K	R	A	C	I	N	U	D	P	T	B
E	O	N	N	P	G	N	P	S	J	F	J	B	Z	N	O	A	E	B	Y
C	O	I	A	A	B	E	H	E	U	K	A	Q	C	B	R	T	M	H	H
Q	R	H	Z	K	A	G	T	D	A	I	H	T	Y	G	D	Y	C	R	D
B	S	C	F	K	U	E	I	G	F	Q	I	M	O	K	N	R	U	Y	T
Y	I	T	H	Z	S	N	Z	S	T	O	A	T	X	V	A	C	J	S	W
X	P	A	L	S	M	I	I	S	N	T	P	Z	J	E	T	E	E	Q	Z
W	W	M	S	L	K	T	Q	D	J	Y	U	W	S	G	N	G	K	W	J
B	O	G	W	F	A	N	I	F	R	J	W	Y	F	Q	E	C	X	S	Q
Z	A	N	C	S	J	A	J	C	D	J	R	B	Z	U	M	I	T	N	Q
O	Y	I	N	X	G	Z	Y	F	L	A	E	M	B	X	E	G	C	Y	J
R	W	R	Y	N	W	Y	A	N	N	C	E	U	D	N	C	D	K	C	L
D	K	T	O	H	P	B	B	I	B	M	V	F	B	C	A	D	W	P	G
Q	Q	S	O	U	C	L	B	C	V	L	L	J	L	R	L	N	Q	F	W
C	I	N	Z	P	J	V	E	A	E	L	H	X	Z	P	P	B	D	E	Y
S	O	R	T	I	N	G	N	E	T	W	O	R	K	S	X	J	I	Y	Q

Puzzle generated at:
http://puzzlemaker.school.discovery.com/WordSearchSetupForm.html

Word Search Puzzle

Type 2, With Clues Supplied for the Words to be Found

Seven birds

□ □ □ □ □ □ □

Five units of length

□ □ □ □ □

Four currencies

□ □ □ □

Two things football players wear

□ □

Large gland in the neck

□

L	E	A	G	L	E	U	R	O	K	R	D
P	O	X	W	Y	A	R	D	R	X	E	O
I	E	O	T	H	Y	R	O	I	D	T	L
H	N	S	N	T	E	T	P	B	N	E	L
A	J	C	O	Z	S	L	M	O	I	M	A
W	Z	O	H	C	J	N	M	I	U	N	R
K	F	J	E	R	S	E	Y	E	L	N	B
V	E	G	R	E	T	X	Z	J	T	E	D

USA Today’s “Word Roundup” for May 16, 2007: http://puzzles.usatoday.com/

Q1: Solve the “Word Roundup” puzzle shown above

Q2: Build a “Word Roundup” puzzle where the clue is “15 country names.”

Converting a 2D Search to 1D Searches

L	E	A	G	L	E	U	R	O	K	R	D
P	O	X	W	Y	A	R	D	R	X	E	O
I	E	O	T	H	Y	R	O	I	D	T	L
H	N	S	N	T	E	T	P	B	N	E	L
A	J	C	O	Z	S	L	M	O	I	M	A
W	Z	O	H	C	J	N	M	I	U	N	R
K	F	J	E	R	S	E	Y	E	L	N	B
V	E	G	R	E	T	X	Z	J	T	E	D

LEAGLEUROKRDPOXWYARDRXEOIEOTHYROIDTLHNSNTETPBNEL
AJCOZSLMOIMAWZOHCJNMIUNRKFJERSEYELNBVEGRETXZJTED

LPIHAWKVEOENJZFEAXOSCOJGGWTNOHERLYHTZCREEAYESJST
URRTLNEXRDOPMMYZORIBOIEJKXDNIULTRETEMNNEDOLLARBD

A Challenging Hybrid Word Search Puzzle

E	S	A	H	C	G	T	S	I	H	D	R	E	H	T	O
Y	I	E	M	U	L	E	R	O	W	G	R	P	E	N	A
V	N	H	A	E	B	D	A	T	O	R	L	E	E	F	N
A	E	R	S	L	A	N	E	A	H	A	A	E	A	G	I
R	D	N	I	U	I	N	G	S	N	T	R	T	R	M	M
G	Z	N	A	A	S	E	S	K	D	G	O	E	H	P	E
C	G	L	R	R	N	I	A	R	T	R	E	L	R	D	A
R	M	D	A	J	C	S	A	Y	B	D	A	I	S	O	S
A	O	N	I	O	N	C	C	D	O	R	D	O	A	S	O
P	E	P	O	C	H	H	C	R	R	E	M	A	B	S	A
S	H	C	A	E	P	M	I	E	O	D	A	I	R	T	H
K	D	I	M	I	T	O	W	N	N	W	I	N	O	S	C

CLUES

'*House, M.D.*'s Robert
"It makes its own ____"
Dice game
Basketball position
Google "The"
Unit of time & space
Fuzzy fruit
Shy

Japanese food
"____ to an end"
Whooping bird
Rapper Gold
Water goes down
Wacky witches
'*White and* ____'
The Fifth Element

Sawdust quiets
Infintesimals
Deadly sin
Deadly sin
Deadly sin
Deadly sin
'*Runaway* ____'
Two-by-four

"Walk the ____"
Dolce ____
Drunks
Ethan Rayne
Japanese toons
Peter Gabriel's '*Curtains*'
1.This 2.That 3. ____
'*The Italian Job*'s Seth

Word-Search Puzzle with a Twist

ARIAS
BRANDT
BRIAND
DUNANT
HULL
HUME
KING
MANDELA
MARSHALL
MONETA
MOTT
PAULING
PEARSON
PERES
RABIN
RENAULT
SADAT
SAKHAROV
WALESA

U	H	A	L	E	D	N	A	M	H
T	U	S	A	D	A	T	V	U	N
L	L	E	U	P	S	O	M	E	S
U	L	L					S	A	D
A	D	A					I	L	A
N	U	W					B	T	K
E	N	K					E	I	I
R	A	A	E	O	R	N	I	T	N
S	N	T	E	N	O	A	D	N	G
D	T	T	O	M	R	S	M	T	G

This “Missing Peace Puzzle” was used in a qualifying round of World Puzzle Championships.

Supply the 16 missing letters at the center of the grid so that the word-search puzzle contains 18 of the 19 names of Nobel Peace Prize winners listed.

Q3: Solve the word-search puzzle above and supply the missing name.

String Matching: Problem Definition

Given a data string with n symbols and a pattern string with m symbols:

1. Does the pattern string appear in the data string?

2. What are the locations of all occurrences of the pattern in the data?

The brute-force, or sliding window, algorithm
Consider all possible positions where the pattern might begin ($n - m + 1$)
For each start position, do up to m comparison to see if there is a match

Worst-case complexity = O(mn); e.g., pattern "aaaaa", data "aaaaaaaaaa"

Pattern string of length m = 5 symbols: **EAGLE**

Data string of length n = 96 symbols

EAGLE

```
           12          24          36
LEAGLEUROKRDPOXWYARDRXEOIEOTHYROIDTLHNSNTETPBNEL
AJCOZSLMOIMAWZOHCJNMIUNRKFJERSEYELNBVEGRETXZJTED
48          60          72          84
```

Converting 2D Search Puzzles to 1D Searches

A 2D word search puzzle looks more exotic but it can be readily converted to a 1D string search puzzle

L	E	A	G	L	E	U	R	O	E	X	T
R	A	X	W	Y	A	R	D	R	X	E	O
I	E	O	T	H	Y	R	O	I	D	T	L
H	N	S	N	T	E	T	P	B	N	E	L
A	J	C	O	Z	S	L	M	O	I	M	A
W	Z	O	H	C	J	N	M	I	U	N	R
K	F	J	E	R	S	E	Y	E	L	N	B
V	E	G	R	E	T	X	Z	J	T	E	D

Row-major order

LEAGLEUROEXTRAXWYARDRXEOIEOTHYROIDTLHNSNTETPBNEL
AJCOZSLMOIMAWZOHCJNMIUNRKFJERSEYELNBVEGRETXZJTED

Insert a special symbol (#) between rows to ensure that new words or patterns are not created by the expansion

LEAGLEUROEXT#RAXWYARDRXEO#IEOTHYROIDTL#HNSNTETPBNEL#
AJCOZSLMOIMA#WZOHCJNMIUNR#KFJERSEYELNB#VEGRETXZJTED#

Column-major order

Similarly for (anti)diagonal

Finding a Needle in a Haystack

Search for the 10-symbol "needle" h-e-l-e-n- -h-u-n-t in the Internet "haystack" with many TBs of data

The brute force algorithm amounts to: "Look in this corner, now in this other corner, then over there, and so on."

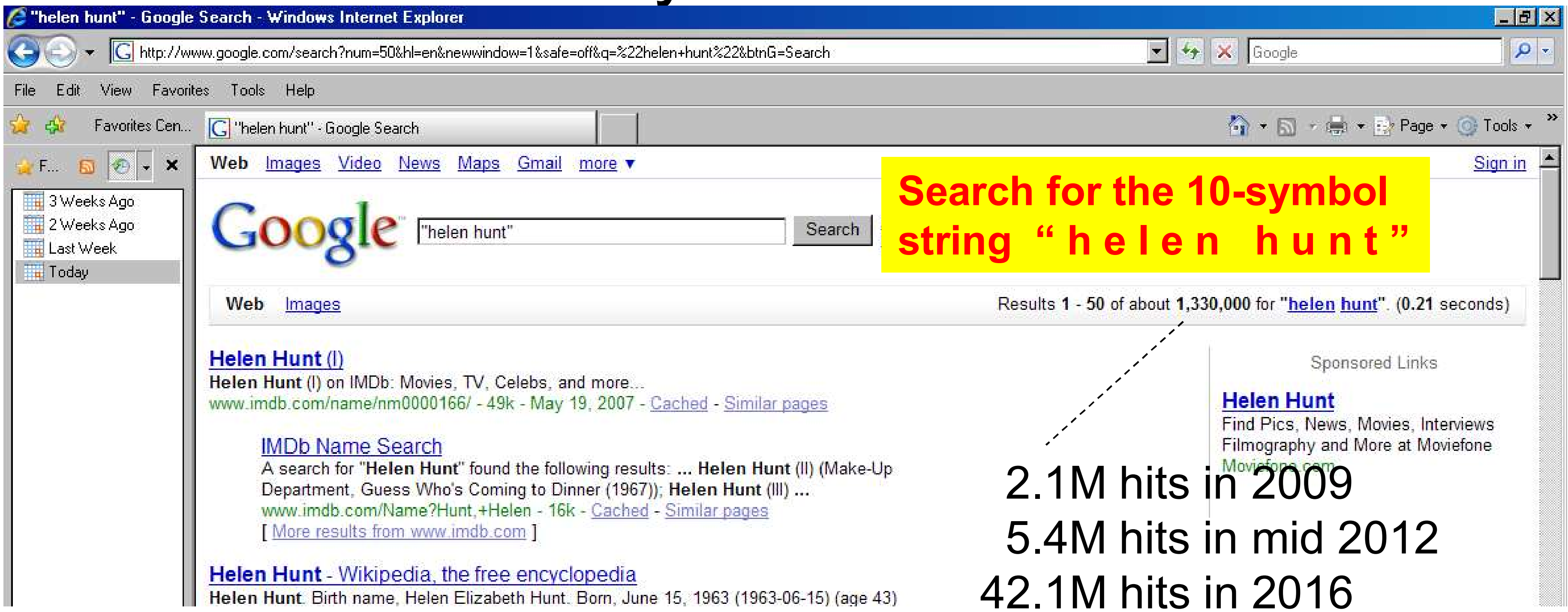
The Internet holds some 1^+ trillion pages, growing by billions a day; each page on average contains in excess of 10 KB, say

$m \cong 10$

$n \cong 10^{12} \times 10^4 = 10^{16}$ B = 10 PB (Petabyte) = 0.01 EB (Exabyte)

$O(mn) \cong 10^{17}$ comparisons $\Rightarrow 10^8$ s (> 3 years), with 10^9 comparisons/s

Needle in a Haystack: Internet Search

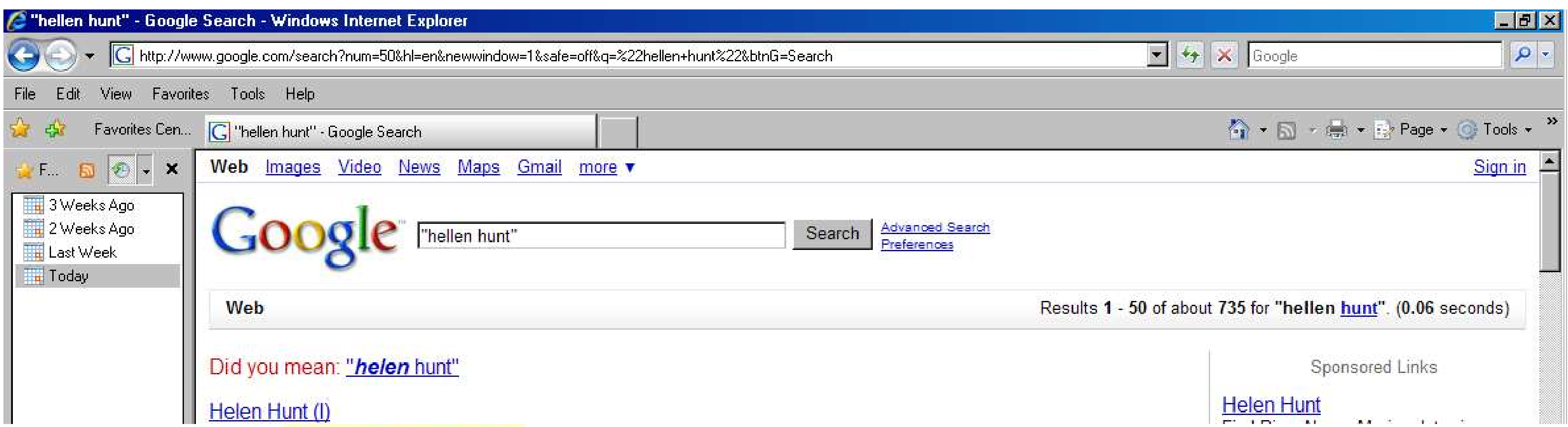


Search for the 10-symbol string "h e l e n   h u n t"

2.1M hits in 2009

5.4M hits in mid 2012

42.1M hits in 2016



Needle in a Haystack: Doing Less Work

For a particular pattern and unpredictable data strings, preprocess the pattern so that searching for it in various data strings becomes faster

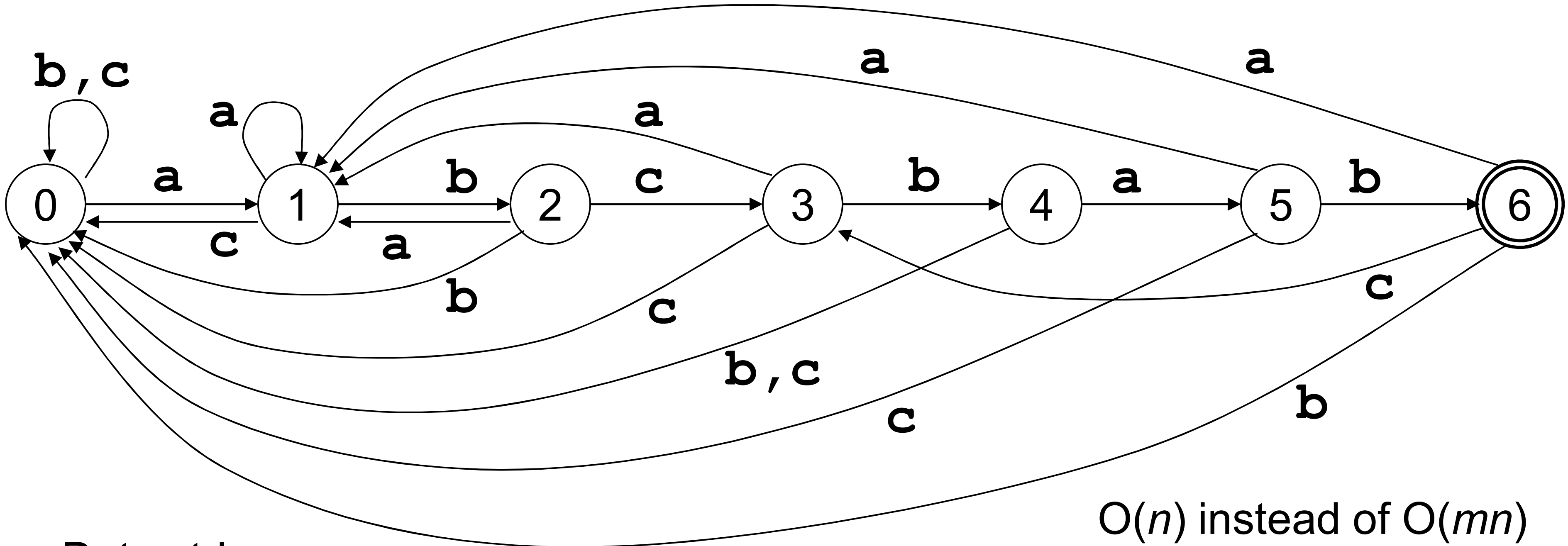
Analogy: Magnetize the needle

For a particular data string and unpredictable patterns, preprocess the data string so that when a pattern is supplied, we can readily find it with much less work

Analogy: Do a thorough search of the haystack for different types of needles and place markers to guide future searches

Example of Preprocessing the Pattern String

Devise an efficient method for finding the pattern “**`abcbab`**” in various data strings formed from the symbols **`a`**, **`b`** and **`c`**



O(n) instead of O(mn)

Data string:

`a b c b b b a b c b a b b c a a b c b a b c b a b c b b`

0 1 2 3 4 0 0 1 2 3 4 5 6 0 0 1 1 2 3 4 5 6 3 4 5 6 3 4 0

Example of Preprocessing the Data String

Devise an efficient method for finding various patterns in the data string:

a	b	c	b	b	b	a	b	c	b	a	b	b	c	a	a	b	c	b	a	b	c	b	a	b	c	b	b
0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27

a a b	14
a b b	10
a b c	0, 6, 15, 19, 23
b a b	5, 9, 18, 22
b b a	4
b b b	3
b b c	11
b c a	12
b c b	1, 7, 16, 20, 24
c a a	13
c b a	8, 17, 21
c b b	2, 25

Find all occurrences of the pattern "**abcbab**"

$\boxed{\text{a b c}}$	0, 6, 15, 19, 23
b c b	1, 7, 16, 20, 24
c b a	8, 17, 21
$\boxed{\text{b a b}}$	5, 9, 18, 22

a b c	0, 6, 15, 19, 23
b c b	1, 7, 16, 20, 24
c b a	8, 17, 21
b a b	5, 9, 18, 22

Q4: Use the index above to find the locations of **b a b c** in the string.

Another Preprocessing Example: Suffix Tree

A suffix tree is a tree in which root-to-leaf paths correspond to all the suffixes of a string

Review article: *CACM*, April 2016, pp. 66-73.

Allows searching for a substring of length *m* in *O*(*m*) time instead of *O*(*n*) time.

Example: Does the string "*mississippi$*" contain the substring "*siss*"? What about "*pie*"?

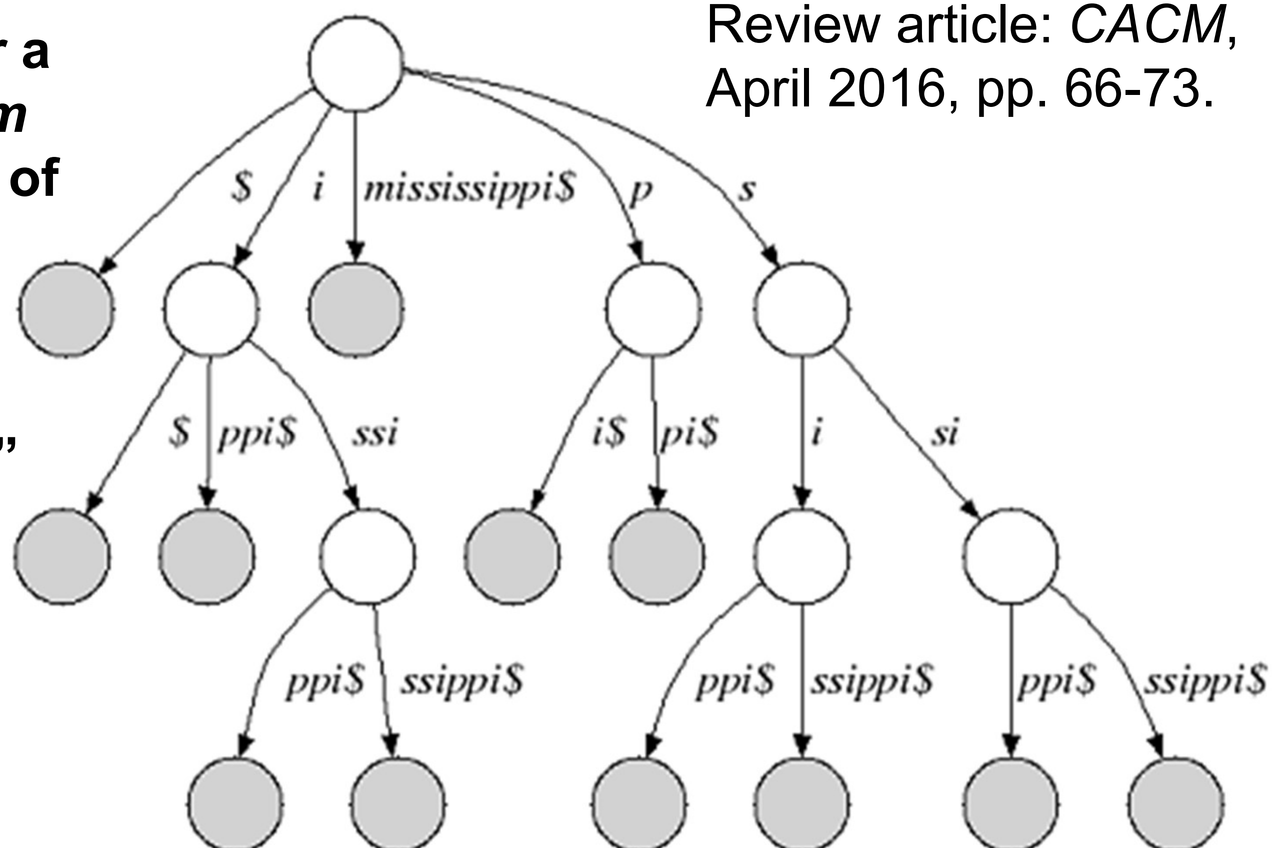


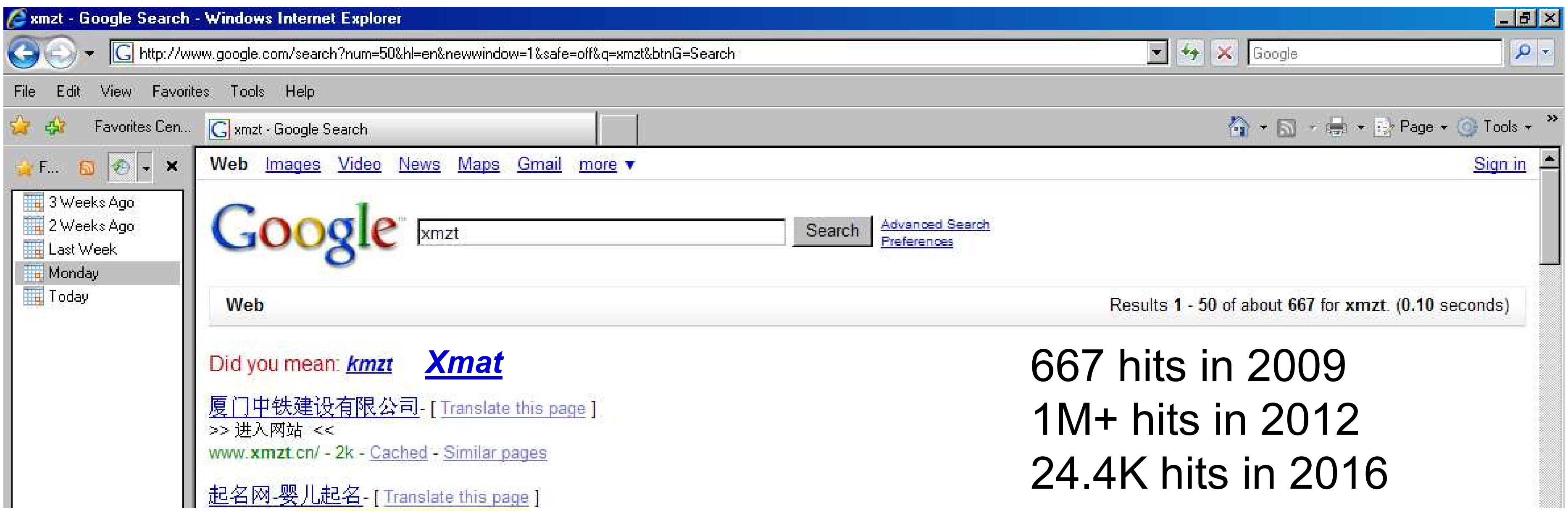
Image source:
http://docs.seqan.de/seqan/1.2/streeSentinel.png

Search Engine Indexes

17.5B hits in 2009

25.3B hits in 2012

25.3B hits in 2016



667 hits in 2009

1M+ hits in 2012

24.4K hits in 2016

Approximate String Matching

Notion of string distance
Each of the following transformations in a string creates a distance of 1

1. Insertion of an extra symbol
2. Deletion of a symbol
3. Transposition of two adjacent symbols

Example distance-1 strings for **`helen hunt:`**

`hellen hunt`	Insertion
`elen hunt`	Deletion
`helen hnut`	Transposition

Example distance-2 strings for **`helen hunt:`**

`hellen hnut`	Insertion + Transposition
`elen huntt`	Deletion + Insertion
`lheen hunt`	2 Transpositions

Wildcard symbols can help in formulating approximate string searches

`h* hunt` means any string that begins with an "**`h`**", ends with "**`hunt`**", and has an arbitrary set of symbols between the two

Melvyl (UC library catalog) allows such searches, e.g., author: **`hunt h*`**

The (DNA) Sequence Alignment Problem

Given sequences S1 and S2 composed of the letters A, C, G, T
Determine their degree of similarity

S1: **A G G G C T**
S2: **A G G C A**

Application: Matching a given DNA sequence
to a set of sequences in a database to find the best match

Dissimilarity arises from missing letters or mismatched letters

Alignment 1:

S1: **A G G G C T**
S2: **A G G C A –**

Penalty = GC mismatch
+ CA mismatch
+ 1 gap

Alignment 2:

S1: **A G G G C T**
S2: **A G G – C A**

Penalty = 1 gap
+ TA mismatch

Optimal alignment found via dynamic programming

A Couple of Bonus Word Search Puzzles

```
z r g r d r f g m l s f z b k s o w w i x k x f r r e s l
a i l a w r n s y x r s h q e t p d h t s i d d a v o c a
e x e h g u k u h f x x j w y x w z h r i h p e e v e p r
f c n g w x v j l z q q m a v s w m d f t k h e t m l g p
d d w m u y s l b q j u j d m m r w p g v y q k c y j d y
j z o u e m z g e s j i o y f k w d x y a h g c c u c o a
q d x q m f j v x c a w i q u v u s m s d s w s z n l u t
j w p s p q f y o i r n u j l n a v n t q w h p t o o b c
n h u q l f r w o a d z u l b t a o i p y v d f h c r l s
l n e w v z j z w f b j w u a o c t a x n g z f w b f e r
s u d e l q n p i m z p n l m t y l j r p l l h g y h n i
j v j i g n p k e m w q n d j q m w i c f o z f f v f h c
r i s i a i u u t s d i y v o h m a a m z o m o i u x j o
h i g k n n e g b f j t g u c c z n e t h e u v p t r y l
s t u t e w m p u z z l e r s n a l w a y s c t f o l d h
t m k w h i l z c u y j j w h s x n r a q n g e z s i k t
w l l h a n v s f s a p c d d r a i m e t g r w w g h f z
a h r s p z v s d f r t h v i e t y e a z i c w a h j b f
e n d y e o b a r o n n x k v b o l d f x f t a b o o k v
t w c n u d l l o n i v v t j c s i o b p m n h i h d p i
v f t r a b u u c h q p c u p p m p l a u t y f q r l i z
l y w f o z y o a x e t h e q c t d u v h p j o m s k g g
a s d l m n u s i c x u t d a i w v e z z y j e h p l d q
a h h n w m n n f u d y k a e b a v l b z i b e r b z k o
b u t q o i d i u i v y n a k s s g p n f w p d s f e i x
c r e a s e i m k c r o s s m h c l o u d e r b p c l u e
t h n j k k l i u f v m t o e i z x u c g n b p s y b x x
s e r s d j g e a n c s d o f u h o l t c r x i j f i m z
g t n n l u p n h f j k a c z u u t v x o j h w c w a j l
```

ASTUTE	ALWAYS	BEND	BARON
BOLD	BOOK	CREASE	CROSS
CLOUD	CLUE	DOUBLE	EAR
EVE	FIT	FUR	GAG
GIG	GNU	GUN	HAIL
IRIS	JOCKEY