MA/CSSE 473 Day 27

Boyer-Moore

Hashing

MA/CSSE 473 Day 27

Assignment	Old due date	New due date
10	Tuesday, Oct 19	Wednesday, Oct 20
Convex Hull	Thursday, Oct 21	Friday, Oct 22
11	Saturday, Oct 23	Tuesday, Oct 26
12	Tuesday, Oct 26	Thursday, Oct 28
13	Thursday, Oct 28	Wednesday, Nov 3

Tomorrow!

- **Take-home exam** available by Oct 29 (Friday) at 9:55 AM, due Nov 1 (Monday) at 8 AM.
- **Student Questions**
- Boyer-Moore
- Hashing review

Recap: Boyer-Moore

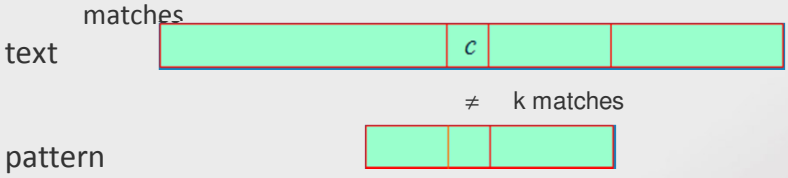
- Boyer-Moore takes into account k, the number of matched characters (from the right) before a mismatch occurs.
- If k=0, we simply do the same shift as Horspool's algorithm.

Boyer-Moore Algorithm

- Based on same two main ideas:
- compare pattern characters to text characters from right to left
- precompute the shift amounts in two tables
  - **bad-symbol table** indicates how much to shift based on the text's character that causes a mismatch
  - **good-suffix table** indicates how much to shift based on matched part (suffix) of the pattern

Bad-symbol shift in Boyer-Moore

- If the rightmost character of the pattern does not match, Boyer-Moore algorithm acts much like Horspool's
- If the rightmost character of the pattern does match, BM compares preceding characters right to left until either
  - all pattern's characters match, or
  - a mismatch on text's character c is encountered after $k > 0$ matches



bad-symbol shift: How much should we shift by?

$d_1 = \max\{t_1(c) - k, 1\}$,
where $t_1(c)$ is the value form the Horspool shift table.

Good-suffix Shift in Boyer-Moore

- Do quiz questions 1-3

Boyer-Moore Algorithm

After matching successfully $0 < k < m$ characters, the algorithm shifts the pattern right by

$$d = \max\{d_1, d_2\}$$

where $d_1 = \max\{t_1(c) - k, 1\}$ is bad-symbol shift

$d_2(k)$ is good-suffix shift

Boyer-Moore Example

```
pattern = abracadabra
text =
abracadabtabradabracadabcadaxbrabbracadabraxxxxxxabracadabracadabra
m = 11,  n = 67
badCharacterTable: a3 b2 r1 a3 c6 x11
GoodSuffixTable: (1,3) (2,10) (3,10) (4,7) (5,7) (6,7) (7,7) (8,7)
(9,7) (10, 7)

abracadabtabradabracadabcadaxbrabbracadabraxxxxxxabracadabracadabra
abracadabra
i =  10     k =  1     t1 =  11     d1 =  10     d2 =  3
abracadabtabradabracadabcadaxbrabbracadabraxxxxxxabracadabracadabra
          abracadabra
i =  20     k =  1     t1 =  6     d1 =  5     d2 =  3
abracadabtabradabracadabcadaxbrabbracadabraxxxxxxabracadabracadabra
               abracadabra
i =  25     k =  1     t1 =  6     d1 =  5     d2 =  3
abracadabtabradabracadabcadaxbrabbracadabraxxxxxxabracadabracadabra
                    abracadabra
i =  30     k =  0     t1 =  1     d1 =  1
```

Boyer-Moore Example

First step is a repeat from the previous slide

```
abracadabtabradabracadabcadaxbrabbracadabraxxxxxxabracadabracadabra
                    abracadabra
i =  30     k =  0     t1 =  1     d1 =  1
abracadabtabradabracadabcadaxbrabbracadabraxxxxxxabracadabracadabra
                     abracadabra
i =  31     k =  3     t1 =  11     d1 =  8     d2 =  10
abracadabtabradabracadabcadaxbrabbracadabraxxxxxxabracadabracadabra
                               abracadabra
i =  41     k =  0     t1 =  1     d1 =  1
abracadabtabradabracadabcadaxbrabbracadabraxxxxxxabracadabracadabra
                                 abracadabra
i =  42     k =  10     t1 =  2     d1 =  1     d2 =  7
abracadabtabradabracadabcadaxbrabbracadabraxxxxxxabracadabracadabra
                                       abracadabra
i =  49     k =  1     t1 =  11     d1 =  10     d2 =  3
abracadabtabradabracadabcadaxbrabbracadabraxxxxxxabracadabracadabra
                                                  abracadabra
49
```

Horspool took 13 times through the outer loop; Boyer-Moore took 9

On-line Boyer-Moore Example

- On Moore's home page
- http://www.cs.utexas.edu/users/moore/best-ideas/string-searching/fstrpos-example.html

Hashing Review

(Excellent detailed reference: Weiss Chapter 20)

Discuss the following questions in a group of three students

- What problem do we try to solve by hashing?
- What is the brute force approach to solving that problem?
- What alternatives (other than hashing) have we seen?
- What is the general idea of how hashing works?
- Why does it fit into Chapter 7 (space-time tradeoffs)?
- What are the main issues to be addressed when discussing hashing implementation?
- How to choose between a hash table and a binary search tree?

A brief summary of hashing with a nice animation:
http://www.cs.auckland.ac.nz/software/AlgAnim/hash_tables.html

Terminology and analysis

- collision
- load factor (λ)
- perfect hash function
- open addressing
  - linear probing
  - cluster
  - quadratic probing
  - rehashing
- separate chaining

- expected lookup time (as a function of λ)
  - For successful search
  - For unsuccessful search

Q1-10

Some Hashing Details

- The next slides are from CSSE 230.
- They are here in case you didn't "get it" the first time.
- We will not go over all of them in detail in class.
- If you don't understand the effect of clustering, you might find the animation that is linked from the slides especially helpful.

Collision Resolution: Linear Probing

- When an item hashes to a table location occupied by a non-equal item, simply use the next available space.
- Try H+1, H+2, H+3, …
  - With wraparound at the end of the array
- Problem: Clustering (picture on next slide)
- http://www.cs.auckland.ac.nz/software/AlgAnim/hash_tables.html
- We'll let it keep running while we look at analysis.

Figure 20.4
Linear probing hash table after each insertion

```
hash ( 89, 10 ) = 9
hash ( 18, 10 ) = 8
hash ( 49, 10 ) = 9
hash ( 58, 10 ) = 8
hash (  9, 10 ) = 9
```

	After insert 89	After insert 18	After insert 49	After insert 58	After insert 9
0			49	49	49
1				58	58
2					9
3					
4					
5					
6					
7					
8		18	18	18	18
9	89	89	89	89	89

Analysis of linear probing

- Dependent on the **load factor**, λ, which is the ratio of the number of items in the table to the size of the table. Thus $0 \leq \lambda \leq 1$.
- For a given λ, what is the expected number of probes before an empty location is found?
- For simplicity, assume that all locations are equally likely to be occupied, and equally likely to be the next one we look at. Then the probability that a given cell is empty is $1 - \lambda$, and thus the expected number of probes before finding an empty cell is (write it as a summation).

```
> simplify(sum(i*(1-lambda)*lambda^(i-1), i=1..infinity));
```

$$-\frac{1}{\lambda - 1}$$

Analysis of linear probing (continued)

- The "equally likely" probability is not realistic, because of **clustering**
- Large blocks of consecutive occupied cells are formed. Any attempt to place a new item in any of those cells results in extending the cluster by at least one item
- Thus items collide not only because of identical hash values, but also because of hash values that happen to put them into the cluster
- Average number of probes when λ is large:
  - $0.5 [1 + 1/(1- \lambda)^2]$.
    - For a proof, see Knuth, *The Art of Computer Programming*, Vol 3: Searching Sorting, 2nd ed, Addision-Wesley, Reading, MA, 1998.
  - What are the values for λ = 0, 0.5, 0.75, 0.9?
  - **When λ approaches 1, this gets bad!**
  - But if λ is close to zero, then the average is near 1.0

So why consider linear probing?

- Easy to implement
- Simple code has fast run time per probe
- Works well when load factor is low
  - It could be more efficient just to rehash using a bigger table once it starts to fill.
  - What is oftern done in practice: rehash to an array that is double in size once the load factor reaches 0.5
- What about other fast, easy-to-implement strategies?

Quadratic probing

- With linear probing, if there is a collision at H, we try H, H+1, H+2, H+3,... until we find an empty spot.
  - Causes (primary) clustering
- With quadratic probing, we try H, $H+1^2$. $H+2^2$, $H+3^2$,...
  - Eliminates primary clustering, but can cause secondary clustering.

Hints for quadratic probing

- **Choose a prime number for the array size**
  - If the array used for the table is not more than half full, finding a place to do the insertion is guaranteed , and no cell is probed twice
  - Suppose the array size is P, a prime number greater than 3
  - Show by contradiction that if i and j are $\leq \lfloor P/2 \rfloor$, and $i \neq j$, then $H + i^2 \pmod P \not\equiv H + j^2 \pmod P$.
- **Use an algebraic trick to calculate next index**
  - Replaces mod and general multiplication with subtraction and a bit shift
  - Difference between successive probes:
    - $H + (i+1)^2 = H + i^2 + (2i+1)$ [can use bit-shift for the multiplication].
    - nextProbe = nextProbe + (2i+1);
      if (nextProbe >= P) nextProbe -= P;

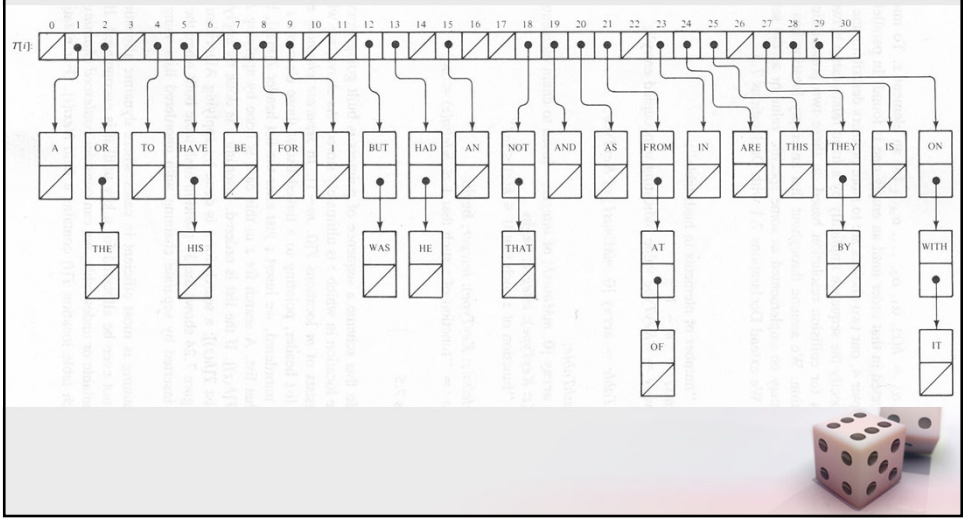
Quadratic probing analysis

- No one has been able to analyze it
- Experimental data shows that it works well
  - Provided that the array size is prime, and is the table is less than half full

Other approaches to collision resolution

- Double hashing
  - A second hash function is used to calculate an offset **d** to use in probing. Try locations h+d, h+2d, h+3d, etc
- Separate chaining
  - Rather than an array of items, we use an array of linked lists. When multiple items hash to the same location, we add them to the list for that location
  - Picture on next slide
  - No clustering effect
    - But we use space (that could have been used to make the array larger) for the links.
    - If many items have the same hash code, the chains can become long.

Hashing with Chaining



Analysis: Hashing with Chaining

- With chaining, the load factor may be > 1.
- Assume a hash function that distributes keys evenly in the table. If there are n keys in the table (backed by an array of size m), the average chain should be λ elements long
- So it takes constant time to compute the hash function plus λ /2 to search within the chain.
- If $\lambda \approx 1$, this is VERY fast
- But there is the extra space for the pointers, which could have been used to make the table larger if open addressing was used