COT 5407: Introduction to Algorithms

Giri Narasimhan

ECS 254A; Phone: x3748

giri@cis.fiu.edu

http://www.cis.fiu.edu/~giri/teach/5407S17.html

https://moodle.cis.fiu.edu/v3.1/course/view.php?id=1494

Room Scheduling Problem

- Given a set of requests to use a room
  - [0,6], [1,4], [2,13], [3,5], [3,8], [5,7], [5,9], [6,10], [8,11], [8,12], [12,14]
- Schedule largest number of above requests in the room
- Different approaches
  - Try by hand, exhaustive search, improve an initial solution, iterative methods, divide and conquer, greedy methods, etc.
- **Simple Greedy Selection**
  - Sort by start time and pick in "greedy" fashion
  - Does not work. WHY?
    - [0,6], [6,10] is the solution you will end up with.
- **Other greedy strategies**
  - Sort by length of interval
  - Does not work. WHY?

Greedy Algorithms

- Given a set of activities (s_i, f_i), we want to schedule the maximum number of non-overlapping activities.
- GREEDY-ACTIVITY-SELECTOR (s, f)

1. n = length[s]
2. S = $\{a_1\}$
3. i = 1
4. **for** m = 2 **to** n **do**
5.   **if** s_m is not before f_i **then**
6.   S = S U $\{a_m\}$
7.   i = m
8. return S

Why does it work?

- THEOREM

  Let A be a set of activities and let a_1 be the activity with the earliest finish time. Then activity a_1 is in some maximum-sized subset of non-overlapping activities.

- PROOF

  Let S' be a solution that does not contain a_1. Let a'_1 be the activity with the earliest finish time in S'. Then replacing a'_1 by a_1 gives a solution S of the same size.

  Why are we allowed to replace? Why is it of the same size?

Then apply induction! How?

Greedy Algorithms – Huffman Coding

- Huffman Coding Problem

  Example: Release 29.1 of 15-Feb-2005 of TrEMBL Protein Database contains **1,614,107** sequence entries, comprising **505,947,503** amino acids. There are 20 possible amino acids. What is the minimum number of bits to store the compressed database?

  ~2.5 G bits or 300MB.

- How to improve this?
- Information: Frequencies are not the same.

Ala (A) 7.72	Gln (Q) 3.91	Leu (L) 9.56	Ser (S) 6.98
Arg (R) 5.24	Glu (E) 6.54	Lys (K) 5.96	Thr (T) 5.52
Asn (N) 4.28	Gly (G) 6.90	Met (M) 2.36	Trp (W) 1.18
Asp (D) 5.28	His (H) 2.26	Phe (F) 4.06	Tyr (Y) 3.13
Cys (C) 1.60	Ile (I) 5.88	Pro (P) 4.87	Val (V) 6.66

- **Idea:** Use shorter codes for more frequent amino acids and longer codes for less frequent ones.

Huffman Coding

2 million characters in file.

A, C, G, T, N, Y, R, S, M

IDEA 1: Use ASCII Code

Each need at least 8 bits,

Total = 16 M bits = 2 MB

IDEA 2: Use 4-bit Codes

Each need at least 4 bits,

Total = 8 M bits = 1 MB

Percentage Frequencies

IDEA 3: Use Variable Length Codes

A	22	11
T	22	10
C	18	011
G	18	010
N	10	001
Y	5	00011
R	4	00010
S	4	00001
		00000

How to Decode?

Need Unique decoding!

Easy for Ideas 1 & 2.

What about Idea 3?

1101011011100100011000000000110

1101011011100100011000000000110

2 million characters in file.

Length = ?

Expected length = ?

Sum up products of frequency times the code length, i.e.,

(.22x2 + .22x2 + .18x3 + .18x3 + .10x3 + .05x5 + .04x5 + .04x5 + .03x5) x 2 M bits =

3.24 M bits = .4 MB