CS660: Algorithms - Lecture 7

Instructor: Hoa Vu

San Diego State University

Edit distance

- Given two strings A, B (e.g., two DNA sequences).
- The minimum number of character insertions, deletions, and substitutions to transform A to B.
- Example: *FOOD* and *MONEY*. Edit distance is 4.
- Define the table ED where $ED[i, j]$ is the edit distance between $A[1 \ldots i]$ and $B[1 \ldots j]$.
- Initialization:
  - $ED[0, j] = j$ and $ED[j, 0] = 0$. Transforming the empty string to a string of length j requires j insertions.

Edit distance

- Initialization:
  - $ED[0, j] = j$ and $ED[j, 0] = 0$. Transforming the empty string to a string of length j requires j insertions.
- Filling the table:
- For $i = 1$ to m:
  - $ED[i, 0] \leftarrow i$.
  - For $j = 1$ to n:
    - $ins \leftarrow ED[i, j-1] + 1$
    - $del \leftarrow ED[i-1, j] + 1$
    - If $A[i] = B[j]$ then $rep \leftarrow ED[i-1, j-1]$
    - Else, $rep \leftarrow ED[i-1, j-1] + 1$.
    - $ED[i, j] \leftarrow min\{ins, del, rep\}$.
- Read 3.7.

Greedy algorithms

Algorithm that performs each step by some greedy choice.

Storing files on tape

- Have a set of files that we want to store on a magnetic tape.
- Each file i as a length $L[i]$.
- If the file is stored from 1 to n, the cost to access the i file be:

$$cost(i) = \sum_{k=1}^{i} L(k).$$

- If each file is accessed equally likely, the expected cost is

$$E[cost] = \sum_{i=1}^{n} Pr[\text{pick file } i] cost(i) = 1/n \cdot \sum_{i=1}^{n} \sum_{k=1}^{i} L(k)$$

Storing files on tape

- We reorder the file so that $\pi(i)$ is the file at position i.
- Find π that minimizes the expected cost.
- Let $L(1) = 10, L(2) = 2$. Consider two orderings: $1, 2$, and $2, 1$.
- Idea: order the files by increasing length.
- Proof of correctness: why does this algorithm give you the optimal cost?

Storing files on tape

- $E(cost(\pi)) = \sum_{i=1}^{n} \sum_{k=1}^{i} L(\pi(k))$.
- Idea: order the files by increasing length.
- Proof of correctness: why does this algorithm give you the optimal cost?
- Claim: If π is optimal, then for all i, $L(\pi(i)) \leq L(\pi(i+1))$.
- Proof: Suppose π is optimal and there is some i such that $L(\pi(i)) > L(\pi(i+1))$. Swap the oder of the two files $\pi(i)$ and $\pi(i+1)$. The expected cost changes by $(L(b) - L(a))/n < 0$. Hence, we get a lower cost which means π is not optimal which is a contradiction.