6.851 Advanced Data Structures (Spring'07)

Prof. Erik Demaine TA: Oren Weimann

Problem 5 – Solution

Approximate Pattern Matching Under the Edit Distance Metric.

Definition: An e-path in the dynamic program table is a path that starts in row zero and specifies a total of exactly e errors (mismatches, insertions and deletions).

Definition: An e-path is *farthest reaching in diagonal d* if it is an e-path that ends in diagonal d, and the index of its ending column is largest among such e-paths.

To begin, when $e = 0$, the farthest reaching 0-path ending on diagonal d corresponds to the LCP of $P[1..m]$ and $T[d..n]$. For $e > 0$, the farthest reaching e-path on diagonal d can be found by considering the following three paths that end in diagonal d.

- the farthest reaching $(e-1)$-path on diagonal $d+1$, followed by one vertical edge (deletion from P) to diagonal d, followed by the maximal extension along diagonal d that corresponds to identical substrings in P and T.
- the farthest reaching $(e-1)$-path on diagonal $d-1$, followed by one horizontal edge (deletion from T) to diagonal d, followed by the maximal extension along diagonal d that corresponds to identical substrings in P and T.
- the farthest reaching $(e-1)$-path on diagonal d, followed by one diagonal edge (mismatch), followed by the maximal extension along diagonal d that corresponds to identical substrings in P and T.

Notice that each "maximal extension" can be found in O(1) time using LCA queries on a suffix tree of $P\#T$. Therefore, we can compute the value of the farthest reaching k-paths on all diagonals in $O(nk)$ time ($O(n)$ diagonals, k locations on each diagonal). Any k-path that reaches row m in column c say, means that the edit distance between P and a suffix of $T[1..c]$ is at most k.