Applications of Suffix Tree

Let us have a glimpse of the numerous applications of suffix trees.

Exact String Matching

As already mentioned earlier, given the suffix tree of the text, all occ occurrences of a pattern P can be found in time $\mathcal{O}(|P| + occ)$.

Even if we take into account the time for constructing the suffix tree, this is asymptotically as fast as Knuth–Morris–Pratt for a single pattern and Aho–Corasick for multiple patterns.

However, the primary use of suffix trees is in indexed string matching, where we can afford to spend a lot of time in preprocessing the text, but must then answer queries very quickly.

Approximate String Matching

Several approximate string matching algorithms achieving $\mathcal{O}(kn)$ worst case time complexity are based on suffix trees (see exercises for an example).

Filtering algorithms that reduce approximate string matching to exact string matching such as partitioning the pattern into $k+1$ factors, can use suffix trees in the filtering phase.

Another approach is to generate all strings in the k-neighborhood of the pattern, i.e., all strings within edit distance k from the pattern and search for them in the suffix tree.

The best practical algorithms for indexed approximate string matching are hybrids of the last two approaches. For example, partition the pattern into $\ell \leq k+1$ factors and find approximate occurrences of the factors with edit distance $\lfloor k/\ell \rfloor$ using the neighborhood method in the filtering phase.

Text Statistics

Suffix tree is useful for computing all kinds of statistics on the text. For example:

- Every locus in the suffix tree represents a factor of the text and, vice versa, every factor is represented by some locus. Thus the number of distinct factors in the text is exactly the number of distinct locuses, which can be computed by a traversal of the suffix tree in $\mathcal{O}(n)$ time even though the resulting value is typically $\Theta(n^2)$.

- The longest repeating factor of the text is the longest string that occurs at least twice in the text. It is represented by the deepest internal node in the suffix tree.

Generalized Suffix Tree

A generalized suffix tree of two strings S and T is the suffix three of the string $S£T\$$, where £ and \$ are symbols that do not occur elsewhere in S and T.

Each leaf is marked as an S-leaf or a T-leaf according to the starting position of the suffix it represents. Using a depth first traversal, we determine for each internal node if its subtree contains only S-leafs, only T-leafs, or both. The deepest node that contains both represents the longest common factor of S and T. It can be computed in linear time.

The generalized suffix tree can also be defined for more than two strings.

AC Automaton for the Set of Suffixes

As already mentioned, a suffix tree with suffix links is essentially an Aho–Corasick automaton for the set of all suffixes.

- We saw that it is possible to follow suffix link / failure transition from any locus, not just from suffix tree nodes.

- Following such an implicit suffix link may take more than a constant time, but the total time during the scanning of a string with the automaton is linear in the length of the string. This can be shown with a similar argument as in the construction algorithm.

Thus suffix tree is asymptotically as fast to operate as the AC automaton, but needs much less space.

Matching Statistics

The matching statistics of a string $S[0..n)$ with respect to a string T is an array $MS[0..n)$, where $MS[i]$ is a pair (ℓ_i, p_i) such that

1. $S[i..i+\ell_i)$ is the longest prefix of S_i that is a factor of T, and
2. $T[p_i..p_i+\ell_i) = S[i..i+\ell_i)$.

Matching statistics can be computed by using the suffix tree of T as an AC-automaton and scanning S with it.

- If before reading $S[i]$ we are at the locus (v, d) in the automaton, then $S[i-d..i) = T[j..j+d)$, where $j = \mathit{start}(v)$. If reading $S[i]$ causes a failure transition, then $MS[i-d] = (d, j)$.
- Following the failure transition decrements d and thus increments $i-d$ by one. Following a normal transition/edge, increments both i and d by one, and thus $i-d$ stays the same. Thus all entries are computed.

From the matching statistics, we can easily compute the longest common factor of S and T. Because we need the suffix tree only for T, this saves space compared to a generalized suffix tree.

Matching statistics are also used in some approximate string matching algorithms.

LCA Preprocessing

The lowest common ancestor (LCA) of two nodes u and v is the deepest node that is an ancestor of both u and v. Any tree can be preprocessed in linear time so that the LCA of any two nodes can be computed in constant time. The details are omitted here.

- Let w_i and w_j be the leaves of the suffix tree of T that represent the suffixes T_i and T_j. The lowest common ancestor of w_i and w_j represents the longest common prefix of T_i and T_j. Thus

  $$lpc(T_i, T_j) = depth(LCA(w_i, w_j)),$$

  which can be computed in constant time using the suffix tree with LCA preprocessing.

- The longest common prefix of two suffixes S_i and T_j from two different strings S and T is called the longest common extension. Using the generalized suffix tree with LCA preprocessing, the longest common extension for any pair of suffixes can be computed in constant time.

Some $\mathcal{O}(kn)$ worst case time approximate string matching algorithms use longest common extension data structures (see exercises).

Longest Palindrome

A palindrome is a string that is its own reverse. For example, `saippuakauppias` is a palindrome.

We can use the LCA preprocessed generalized suffix tree of a string T and its reverse T^R to find the longest palindrome in T in linear time.

- Let k_i be the length of the longest common extension of T_{i+1} and T^R_{n-i}, which can be computed in constant time. Then $T[i - k_i..i + k_i]$ is the longest odd length palindrome with the middle at i.

- We can find the longest odd length palindrome by computing k_i for all $i \in [0..n)$ in $\mathcal{O}(n)$ time.

- The longest even length palindrome can be found similarly in $\mathcal{O}(n)$ time. The longest palindrome overall is the longer of the two.

Suffix Array

The suffix array of a text T is a lexicographically ordered array of the set $T_{[0..n]}$ of all suffixes of T. More precisely, the suffix array is an array $SA[0..n]$ of integers containing a permutation of the set $[0..n]$ such that $T_{SA[0]} < T_{SA[1]} < \cdots < T_{SA[n]}$.

A related array is the inverse suffix array SA^{-1} which is the inverse permutation, i.e., $SA^{-1}[SA[i]] = i$ for all $i \in [0..n]$. The value $SA^{-1}[j]$ is the lexicographical rank of the suffix T_j

As with suffix trees, it is common to add the end symbol $T[n] = \$$. It has no effect on the suffix array assuming $ is smaller than any other symbol.

Example 4.7: The suffix array and the inverse suffix array of the text $T =$ banana$.

i	$SA[i]$	$T_{SA[i]}$	j	$SA^{-1}[j]$	
0	6	$	0	4	banana$
1	5	a$	1	3	anana$
2	3	ana$	2	6	nana$
3	1	anana$	3	2	ana$
4	0	banana$	4	5	na$
5	4	na$	5	1	a$
6	2	nana$	6	0	$

Suffix array is much simpler data structure than suffix tree. In particular, the type and the size of the alphabet are usually not a concern.

- The size on the suffix array is $\mathcal{O}(n)$ on any alphabet.

- We will later see that the suffix array can be constructed in the same asymptotic time it takes to sort the characters of the text.

Suffix array construction algorithms are quite fast in practice too. Probably the fastest way to construct a suffix tree is to construct a suffix array first and then use it to construct the suffix tree. (We will see how in a moment.)

Suffix arrays are rarely used alone but are augmented with other arrays and data structures depending on the application. We will see some of them in the next slides.

Exact String Matching

As with suffix trees, exact string matching in T can be performed by a prefix search on the suffix array. The answer can be conveniently given as a contiguous interval $SA[b..e)$ that contains the suffixes with the given prefix. The interval can be found using string binary search.

- If we have the additional arrays $LLCP$ and $RLCP$, the result interval can be computed in $\mathcal{O}(|P| + \log n)$ time.

- Without the additional arrays, we have the same time complexity on average but the worst case time complexity is $\mathcal{O}(|P| \log n)$.

- We can then count the number of occurrences in $\mathcal{O}(1)$ time, list all occ occurrences in $\mathcal{O}(occ)$ time, or list a sample of k occurrences in $\mathcal{O}(k)$ time.

We will later see a quite different method for prefix searching called backward search.

LCP Array

Efficient string binary search uses the arrays $LLCP$ and $RLCP$. However, for many applications, the suffix array is augmented with the lcp array of Definition 1.7 (Lecture 2, slide 21). For all $i \in [1..n]$, we store

$$LCP[i] = lcp(T_{SA[i]}, T_{SA[i-1]})$$

Example 4.8: The LCP array for $T =$ `banana$`.

i	$SA[i]$	$LCP[i]$	$T_{SA[i]}$
0	6		`$`
1	5	0	`a$`
2	3	1	`ana$`
3	1	3	`anana$`
4	0	0	`banana$`
5	4	0	`na$`
6	2	2	`nana$`

Using the solution of Exercise 3.1 (construction of compact trie from sorted array and LCP array), the suffix tree can be constructed from the suffix and LCP arrays in linear time.

However, many suffix tree applications can be solved using the suffix and LCP arrays directly. For example:

- The longest repeating factor is marked by the maximum value in the LCP array.

- The number of distinct factors can be compute by the formula

$$\frac{n(n+1)}{2} + 1 - \sum_{i=1}^{n} LCP[i]$$

  since it equals the number of nodes in the uncompact suffix trie, for which we can use Theorem 1.9.

- Matching statistics of S with respect to T can be computed in linear time using the generalized suffix array of S and T (i.e., the suffix array of $S£T\$$) and its LCP array (exercise).