Substring

From Wikipedia, the free encyclopedia

A **subsequence**, **substring**, **prefix** or **suffix** of a string is a subset of the symbols in a string, where the order of the elements is preserved. In this context, the terms *string* and *sequence* have the same meaning.

Contents

- 1 Subsequence
- 2 Substring
- 3 Prefix
- 4 Suffix
- 5 Border
- 6 Superstring
- 7 References

Subsequence

Main article subsequence

A subsequence of a string $T = t_1 t_2 \ldots t_n$ is a string $\hat{T} = t_{i_1} \ldots t_{i_m}$ such that $i_1 < \ldots < i_m$, where $m \leq n$. Subsequence is a generalisation of substring, suffix and prefix. Finding the longest string which is equal to a subsequence of two or more strings is known as the longest common subsequence problem.

Example: The string anna is equal to a subsequence of the string banana:

```
banana
 || ||
 an na
```

Including the empty subsequence, the number of subsequences of a string of length n where symbols only occur once, is simply the number of subsets of the symbols in the string, i.e. 2^n.

Substring

A substring (or factor) of a string $T = t_1 \ldots t_n$ is a string $\hat{T} = t_{1+i} \ldots t_{m+i}$, where $0 \leq i$ and $m + i \leq n$. A substring of a string is a prefix of a suffix of the string, and equivalently a suffix of a prefix. If $\hat{T}$ is a substring of T, it is also a subsequence, which is a more general concept. Given a pattern P, you can find its occurrences in a string T with a string searching algorithm. Finding the longest string which is equal to a substring of two or more strings is known as the longest common substring problem.

Example: The string ana is equal to substrings (and subsequences) of banana at two different offsets:

```
banana
```

```
 |||||
ana||
  |||
  ana
```

In the mathematical literature, substrings are also called **subwords** (in America) or **factors** (in Europe).

Not including the empty substring, the number of substrings of a string of length n where symbols only occur once, is the number of ways to choose two distinct places between symbols to start/end the substring. Including the very beginning and very end of the string, there are $n + 1$ such places. So there are $\binom{n+1}{2} = \frac{n(n+1)}{2}$ non-empty substrings.

Prefix

A prefix of a string $T = t_1 \dots t_n$ is a string $\widehat{T} = t_1 \dots t_m$, where $m \leq n$. A *proper prefix* of a string is not equal to the string itself ($0 \leq m < n$);[1] some sources[2] in addition restrict a proper prefix to be non-empty ($0 < m < n$). A prefix can be seen as a special case of a substring.

Example: The string ban is equal to a prefix (and substring and subsequence) of the string banana:

```
banana
|||
ban
```

The square subset symbol is sometimes used to indicate a prefix, so that $\widehat{T} \sqsubseteq T$ denotes that $\widehat{T}$ is a prefix of T. This defines a binary relation on strings, called the prefix relation.

In formal language theory, the term *prefix of a string* is also commonly understood to be the set of all prefixes of a string, with respect to that language. See the article on string functions for more details.

Suffix

A suffix of a string $T = t_1 \dots t_n$ is a string $\widehat{T} = t_{n-m+1} \dots t_n$, where $m \leq n$. A *proper suffix* of a string is not equal to the string itself ($0 < m \leq n$); again, a more restricted interpretation is that it is also not empty[1] ($0 < m < n$). A suffix can be seen as a special case of a substring.

Example: The string nana is equal to a suffix (and substring and subsequence) of the string banana:

```
banana
  ||||
  nana
```

Border

A border is suffix and prefix of the same string, e.g. "bab" is a border of "babab".

Superstring

Given a set of k strings $P = \{s_1, s_2, s_3, \ldots s_k\}$, a **superstring** of the set P is single string that contains every string in P as a substring. For example, a concatenation of the strings of P in any order gives a trivial superstring of P. For a more interesting example, let P = {abcc,efab,bccla}. Then bcclabccefab is a superstring of P, and efabccla is another, shorter superstring of P. Generally, we are interested in finding superstrings whose length is small.

References

1. ^ Kelley, Dean (1995). *Automata and Formal Languages: An Introduction*. London: Prentice-Hall International. ISBN 0-13-497777-7.
2. ^ Gusfield, Dan (1999) [1997]. *Algorithms on Strings, Trees and Sequences: Computer Science and Computational Biology*. USA: Cambridge University Press. ISBN 0-521-58519-8.