

Búsqueda en cadena de textos

Reconocimiento de cadenas

Una cadena es una determinada secuencia finita de símbolos pertenecientes a un alfabeto, también finito. Por ejemplo, los símbolos que componen el alfabeto alfanumérico, el código ASCII compuesto por 256 símbolos, entre otros.

El problema será reconocer la ocurrencia de una cadena en un texto (también se conoce con el término de *match*, para afirmar que se encontró la cadena buscada).

Algunos conceptos que se utilizarán en los distintos algoritmos:

Prefijo

Dada la cadena x, y, z se dice que x es prefijo de xy .

Sufijo

Dada la cadena x, y, z se dice que x es sufijo de yx .

Longitud de un texto

Cantidad de caracteres que lo conforman, contando las repeticiones.

Subcadena

Se produce al eliminar unos caracteres.

Ventana de búsqueda

En un texto, la búsqueda o comparación de los elementos de la cadena con los elementos del texto se realiza dentro de los límites de la ventana de búsqueda (de longitud generalmente igual a la de la cadena buscada), comparando cada carácter de la cadena en la ventana con el carácter en el texto alineado con éste, y obteniendo, para cada comparación, un resultado booleano, verdadero o falso. Si todos los caracteres dentro de la ventana coinciden con los correspondientes caracteres del texto, se ha obtenido una ocurrencia; es decir, se hallado la cadena en el texto.

Algoritmos para el reconocimiento de cadenas

Luego de presentarse una ocurrencia (todos los caracteres en la ventana coincidieron con el texto) o un fallo (ningún carácter coincidió, o coincidieron hasta cierta distancia), se debe desplazar la ventana para comenzar una nueva búsqueda; con lo cual, se presentan dos situaciones:

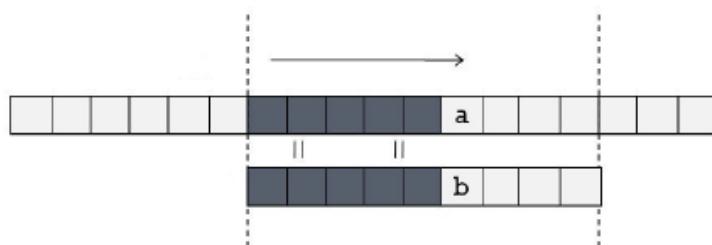
1. Si ha habido una ocurrencia, desplazar la ventana a una distancia igual a la distancia de la cadena. Esto podría implicar que, en el caso de cadenas en las que un sufijo es también prefijo de la misma, se podría perder la posibilidad de encontrarlos en el texto.
2. Si no ha habido ocurrencia, desplazar la ventana un carácter después del carácter fallido. Al igual que en el caso anterior, si en la parte leída que coincide con la cadena existe un sufijo que es también prefijo de ésta, se perdería la posibilidad de encontrar una ocurrencia.

La forma más fácil que permitiría desplazar la ventana de una forma siempre segura es, simplemente, desplazar la ventana una posición, haya o no una ocurrencia. Esto conduce a un **algoritmo de fuerza bruta**, mediante el cual no se perdería ninguna posible ocurrencia, a costa de realizar varias comparaciones sobre los mismos caracteres del texto.

Métodos de búsqueda

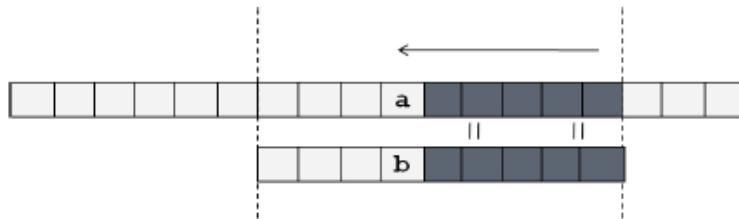
Búsqueda por prefijo

El proceso de comparación de caracteres se realiza de izquierda a derecha dentro de la ventana, buscando el prefijo más largo dentro del texto, que sea, a su vez, prefijo de la cadena:



Búsqueda por sufijo

El proceso de comparación de caracteres se realiza de derecha a izquierda en la ventana, buscando el mayor sufijo, que es también sufijo de la cadena. Esta forma de búsqueda permite, en promedio, evitar leer algunos caracteres del texto:



Algoritmos con búsqueda por prefijo

Knuth-Morris-Pratt (KMP).

Algoritmos con búsqueda por sufijo

Booyer-More (BM).

El algoritmo de búsqueda de cadenas de Knuth-Morris-Pratt o el algoritmo KMP busca las apariciones de un "patrón" dentro de un "texto" principal mediante la observación de que cuando se produce una discrepancia, la palabra en sí contiene información suficiente para determinar dónde podría comenzar la siguiente coincidencia, evitando así el reexamen de caracteres previamente emparejados. El algoritmo fue concebido en 1970 por Donald Knuth y Vaughan Pratt e independientemente por James H. Morris.