Autómata Diccionario

Como se ha visto en prácticas anteriores, una aproximación no determinista al problema del pattern matching permite abordar el problema con mejor comportamiento temporal. En esta práctica se muestra como, a partir del conjunto de patrones a buscar, es posible construir un autómata determinista que permite obtener una solución al problema con menor coste temporal.

El nuevo método se basa en la construcción del *autómata diccionario* del conjunto de patrones. A partir de un conjunto M de palabras sobre determinado alfabeto Σ, se define el autómata diccionario $AD_M = (Q, \Sigma, \delta, q_0, F)$ como sigue:

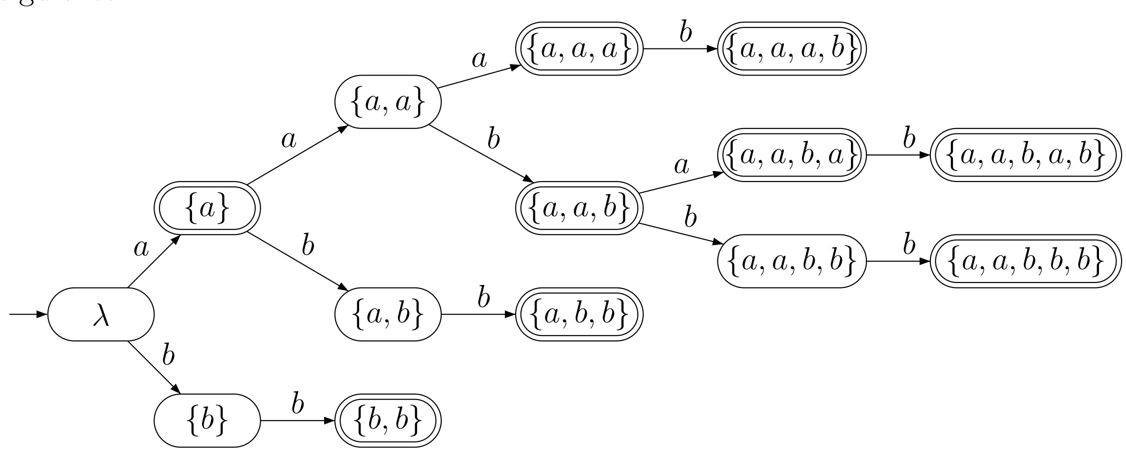
- $Q = \{x \in \Sigma^* \ : \ x \in Pref(M)\}$
- $q_0 = \lambda$
- $F = Pref(M) \cap \Sigma^* M$

  Intuitivamente, el conjunto de estados finales estará formado por los estados que estén identificados con una palabra que tenga, al menos, un sufijo en M.
- $\delta(x, a) = h(xa)$ donde, para cualquier palabra u, $h(u)$ es el sufijo más largo de u que pertenece a $Pref(M)$.

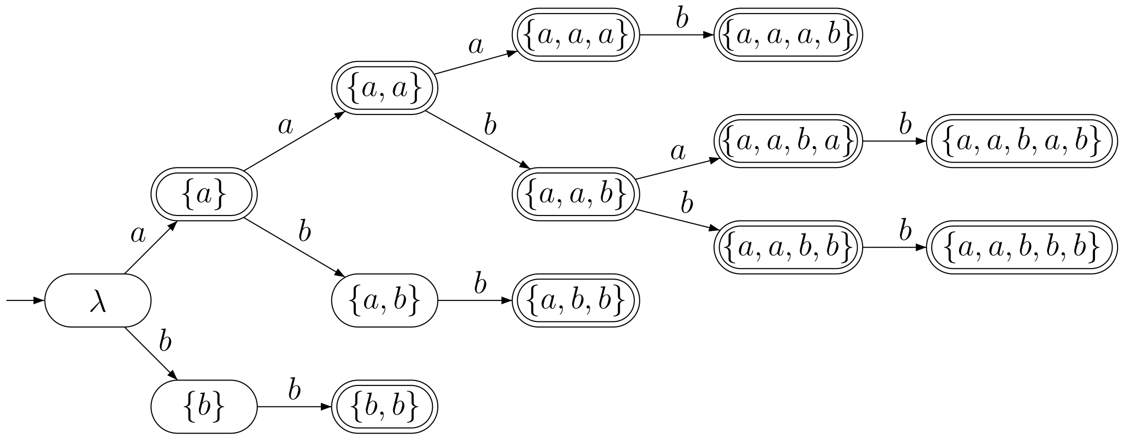
Es importante notar las similitudes entre el autómata diccionario del conjunto M (AD_M) y el árbol aceptor de prefijos del mismo conjunto ($AAP(M)$). Las diferencias residen, por una parte en el conjunto de finales, y por otra parte, en la definición de la función de transición. De hecho, tanto los finales como todas las transiciones de $AAP(M)$ son estados finales y transiciones del AD_M, por lo que en el ejemplo siguiente consideramos el que utilizamos en la práctica anterior:

$$M = \left\{ \begin{array}{l} p_1 = a,\ p_2 = bb,\ p_3 = aaa,\ p_4 = aab,\ p_5 = abb, \\ p_6 = aaab,\ p_7 = aaba,\ p_8 = aabab,\ p_9 = aabbb \end{array} \right\}$$

Posteriormente necesitaremos referirnos a estos patrones individualmente, por lo que hemos asociado a cada patrón un identificador. El árbol aceptor de prefijos correspondiente a este conjunto (donde identificamos cada estado con el prefijo de M con el que está asociado) es el siguiente:

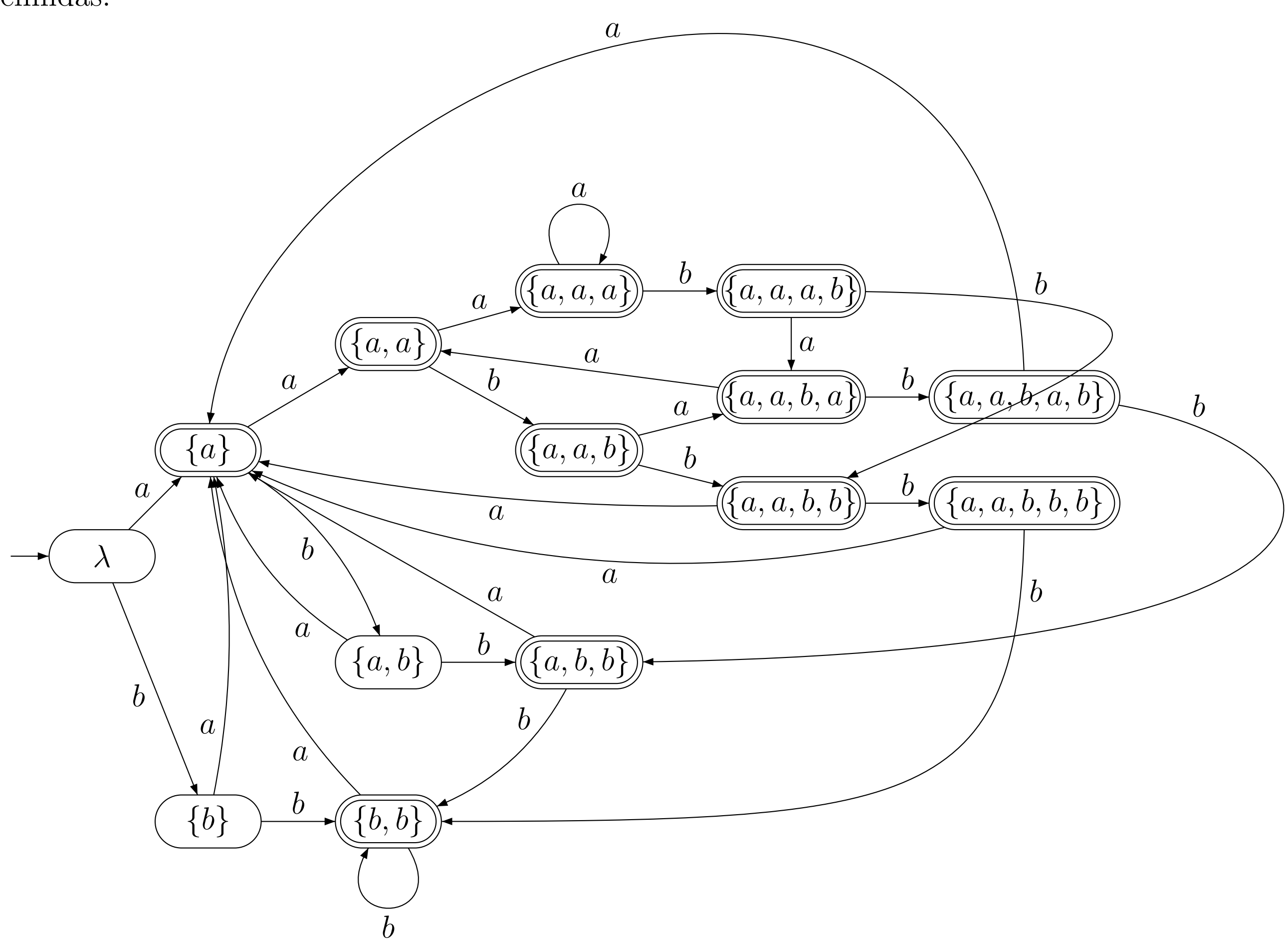


Como primer paso modificamos el conjunto de finales. En ese sentido, nótese como el estado aa, que no es final en el $AAP(M)$, sí es final en AD_M (aa tiene un sufijo que está en M). Esta primera modificación nos lleva al siguiente autómata:



Ahora consideramos las transiciones no definidas en este autómata, entre otras, las transiciones del estado $\{b,b\}$ con los símbolos a y b o del estado $\{a,b\}$ con el símbolo a.

Como ejemplo, aplicando la definición de la función de transición del AD_M, $\delta(\{a,b\},a) = h(\{a,b,a\})$. El sufijo más largo de $\{a,b,a\}$ en $Pref(M)$ es $\{a\}$, por lo tanto $\delta(\{a,b\},a) = \{a\}$. A continuación mostramos el AD_M después de aplicar este proceso a las transiciones no definidas.



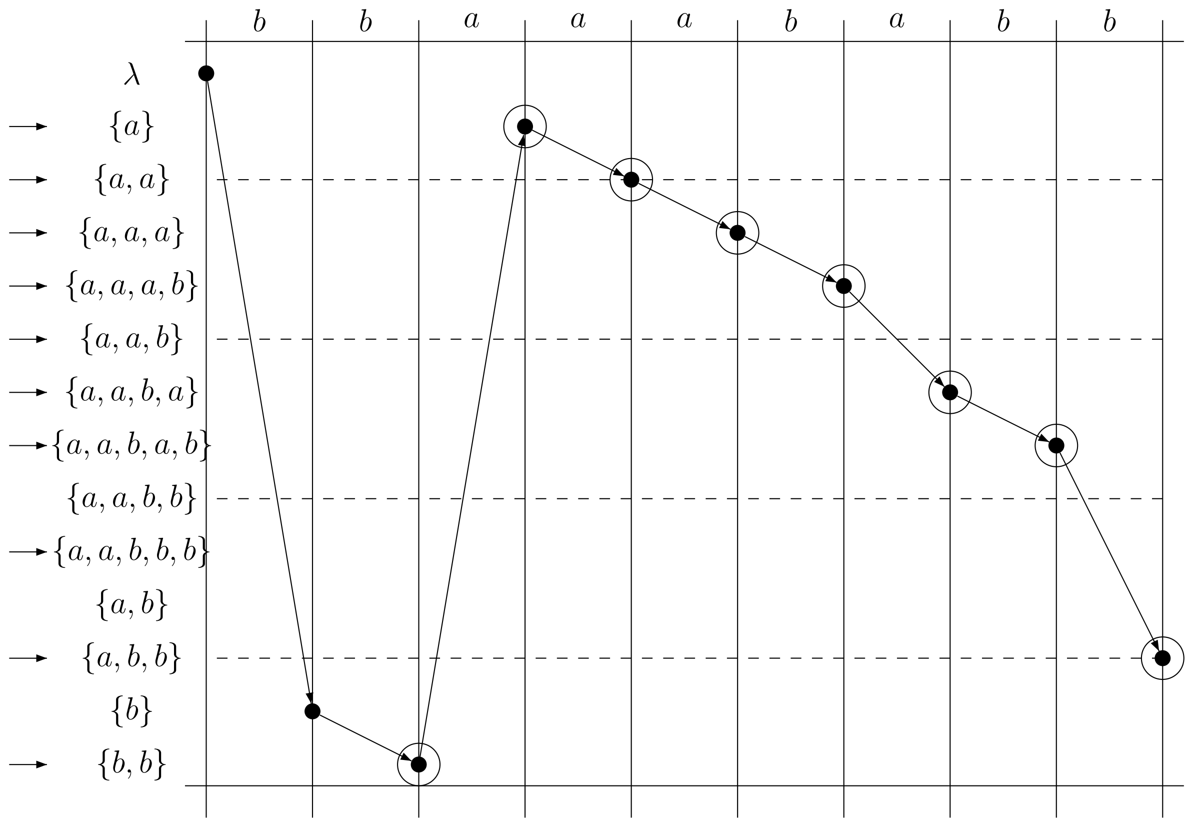
Como sucedía en la práctica anterior, estamos construyendo un autómata (el autómata diccionario A_M) que reconoce el lenguaje $\Sigma^* M$, de modo que, mientras se analiza un texto cualquiera x, alcanzar un estado final implica que se ha encontrado, al menos, un patrón. De hecho, se han encontrado todos los patrones que son sufijo de la palabra que denota el estado final.

Por lo tanto, modificamos el autómata para anotar en cada estado final u, qué patrones son sufijo de la cadena u. Esta información se resume en la siguiente tabla:

estado	$\{a\}$	$\{a,a\}$	$\{b,b\}$	$\{a,a,a\}$	$\{a,a,b\}$	$\{a,b,b\}$
patrones	p_1	p_1	p_2	$p_1,\ p_3$	p_4	$p_2,\ p_5$

estado	$\{a,a,a,b\}$	$\{a,a,b,a\}$	$\{a,a,b,b\}$	$\{a,a,b,a,b$	$\{a,a,b,b,b\}$
patrones	$p_4,\ p_6$	$p_1,\ p_7$	$p_2,\ p_5$	p_8	p_2, p_9

Una vez obtenido el autómata diccionario, es posible detectar todas las posiciones donde aparece una palabra de M en un texto x. Para ello basta realizar un análisis determinista, y siempre que se alcance un estado final, indicar que se han detectado los patrones asociados a los identificadores almacenados en dicho estado final. Por ejemplo, considerando el texto $x = \{b,b,a,a,a,b,a,b,b\}$, el análisis determinista puede representarse como sigue:



En este diagrama hemos marcado los estados finales visitados durante el análisis. Puede verse que después de analizar el segundo símbolo se alcanza el estado $\{b,b\}$, que al ser final indica que se ha detectado un patrón p_2 del conjunto M (el patrón bb). Del mismo modo, por ejemplo: después de analizar $\{b, b, a\}$ y $\{b, b, a, a, a, b, a\}$ se alcanza el estado $\{a\}$ que indica que se ha detectado el patrón a; cuando se ha analizado $\{b, b, a, a, a\}$ se alcanza el estado $\{a, a, a\}$ que indica que se han detectado los patrones a y aaa, y así sucesivamente.

Ejercicios

Ejercicio 1

Implementar un módulo Mathematica que, tomando una palabra u y conjunto de palabras M como entrada, devuelva el sufijo más largo de u que sea un elemento de M.

Ejercicio 2

Implementar un módulo Mathematica que, tomando un conjunto de palabras M como entrada, devuelva el autómata diccionario de ese conjunto.

Ejercicio 3

Implementar un módulo Mathematica para, dados el autómata diccionario de un conjunto de patrones M y un texto x, devuelva el conjunto de posiciones de x en las que aparece un elemento de M.

Bibliografía

Maxime Crochemore, Christophe Hancart and Thierry Lecroq ALGORITHMS ON STRINGS. *Cambridge University Press*, 2007.