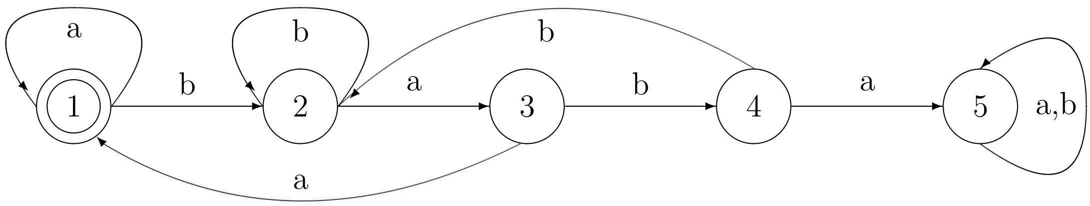
1 Czym będziemy się zajmować

Będziemy konstruować wyrażenie regularne reprezentujące język nad alfabetem z dwoma literami a i b złożony z wszystkich słów, które nie zawierają wzorca $baba$.

Najprostsza metoda rozwiązania tego zadania polega na wykorzystywaniu sposobu tworzenia wyrażeń regularnych zastosowanego w dowodzie twierdzenia o istnieniu takich wyrażeń reprezentujących dowolny język akceptowany przez automat skończony. Przyjrzyjmy się więc odpowiedniemu automatowi.

2 Automat badający występowanie wzorca $baba$



3 Konstrukcja wyrażenia regulanego dla danego automatu skończonego

Mając dany automat skończony o stanach $q_1, q_2, \ldots, q_n$ zwykle rozważa się pomocnicze języki

$$L_{i,j} = \{w \in \Sigma^* : \ \widehat{\delta}(q_i, w) = q_j\}.$$

Jest oczywiste, że są to języki regulane i język L akceptowany przez rozważany automat (podobnie jak każdy inny język akceptowany przez automat skończony) jest sumą mnogościową kilku takich języków. Tak więc znając wyrażenia regularne reprezentujące podane języki łatwo podać wyrażenie regularne reprezentujące język L.

Śledzenie dalszego dowodu może okazać się łatwiejsze, jeżeli diagram automatu skończonego zinterpretujumy jako mapę połączeń drogowych. Na podany diagram możemy patrzeć jak na mapę z pięcioma miastami o numerach od 1 do 5, stany automatu odpowiadają miastom na mapie. Strzałki (krawędzie) automatu wyobrażają drogi. Każda droga jest jednokierunkowa i została nazwana pojedyńczą literą odpowiedniego alfabetu. Jest wyznaczona przez miasto, w którym się zaczyna, i przypisaną jej nazwę. Przy takiej interpretacji diagramów słowa odpowiadają opisom tras przejazdu. Konkretną trasę otrzymujemy wskazując dodatkowo, w jakim mieście zaczyna się. Język $L_{i,j}$ składa się z opisów tych tras, które z miasta (o numerze) i prowadzą do miasta j.

Standardowy dowód wspomnianego twierdzenia wykorzystuje zwykle języki

$$L^k_{i,j} = \{w \in \Sigma^* : \ \widehat{\delta}(q_i, w) = q_j \wedge$$
$$\wedge \forall v \ (v - \text{niepusty, właściwy prefiks } w) \Rightarrow \exists l \leqslant k \ \widehat{\delta}(q_i, v) = q_l\}.$$

Języki $L^k_{i,j}$ zawierają więc wyłącznie opisy tras z $L_{i,j}$ prowadzących przez miasta o numerach $\leqslant k$. Oczywiście, mamy równość $L_{i,j} = L^n_{i,j}$.

Dla tych języków słuszny jest następujący, dość oczywisty wzór:

$$L_{i,j}^{k+1} = L_{i,j}^{k} \ \cup \ L_{i,k+1}^{k}(L_{k+1,k+1}^{k})^{*} L_{k+1,j}^{k}.$$

Wzór ten wynika z tego, że trasy od i do j przez miasta o numerach $\leqslant k+1$ są dwojakiego rodzaju: albo omijają także miasto o numerze $k+1$, albo wymagają przejazdu przez miasto $k+1$, może nawet kilkakrotnego. W tym drugi przypadku składają się z pierwszego przejazdu do $k+1$, z kilku przejazdów z $k+1$ do $k+1$ bez odwiedzania $k+1$ i w końcu z przejazdu z $k+1$ do j.

Przytoczony wzór wyraża najważniejszą część rekurencyjnej definicji języków $L_{i,j}^{k}$. Aby się nim posłużyć wystarczy wyliczyć jeszcze języki $L_{i,j}^{0}$. Zwykle nie sprawia to kłopotów. W przypadku rozważanego automatu mamy na przykład

$$L_{1,1}^{0} = \{\varepsilon, a\}, \ \ L_{2,3}^{0} = \{a\} \ \text{ oraz } \ L_{4,3}^{0} = \emptyset.$$

Języki te bez trudu reprezentujemy wyrażeniami regularnymi, a korzystając z przytoczonego wzoru możemy stopniowo znaleźć wyrażenia regularne reprezentujące wszystkie języki $L_{i,j}^{k}$, a w końcu także wyrażenie regularne reprezentujące L.

4 Uproszczenie podanej metody

Wyliczanie w podany sposób wyrażenia regularnego reprezentującego L jest raczej procesem złożonym. Można go uprościć stosując specyficzną metodę, bardziej dostosowaną do rozpatrywanego automatu, stosując wiele innych wzorów, podobnych do podanego, i korzystając także z innych języków pomocniczych, na przykład języków $L_{i,j}^{X}$ zdefiniowanych dla zbioru liczb naturalnych X wzorem

$$\begin{aligned} L_{i,j}^{X} = \{w \in \Sigma^{*} : \ & \widehat{\delta}(q_i, w) = q_j \wedge \\ & \wedge \forall v \ (v - \text{niepusty, właściwy prefiks } w) \Rightarrow \exists l \notin X \ \widehat{\delta}(q_i, v) = q_l\}, \end{aligned}$$

a więc języków złożonych z opisów tras prowadzących od miasta i do j, omijających miasta o numerach należących do zbioru X.

Nasze zadanie w przypadku narysowanego wyżej automatu wymaga znalezienia wyrażeń regularnych reprezentujących języki $L_{1,1}$, $L_{1,2}$, $L_{1,3}$ i $L_{1,4}$.

Nietrudno zauważyć, że

$$L_{1,4} = L_{1,3}b \ \text{ i podobnie, } \ L_{1,3} = L_{1,2}a.$$

Równie łatwo zauważyć, że zachodzi dość ogólna równość

$$L_{1,2} = L_{1,1}L_{1,2}^{\{1\}}.$$

Każdą trasę z języka $L_{i,j}$ możemy bowiem podzielić na dwa odcinki: pierwszy, w którym być może wielokrotnie przejeżdżamy przez i, kończący się ostatnim przejazdem przez miasto o numerze i, i drugi, zaczynający się w i i prowadzący do j, w którym już nie przejeżdżamy przez i. Podobnie pokazujemy, że

$$L_{1,1} = (L_{1,1}^{\{1\}})^{*}.$$

Z przeprowadzonych rozważań wynika, że rozwiązanie naszego zadania sprowadza się do znalezienia dwóch wyrażeń regularnych ρ_1 i ρ_2, reprezentujących odpowiednio języki $L_{1,1}^{\{1\}}$ oraz $L_{1,2}^{\{1\}}$. Mając te wyrażenia łatwo podać wyrażenie reprezentujące język L, czyli sumę języków $L_{1,1}$, $L_{1,2}$, $L_{1,3}$ i $L_{1,4}$. Jest nim

$$(\rho_1)^{*}(\varepsilon + \rho_2 + \rho_2 a + \rho_2 ab).$$

Zauważmy jeszcze, że

$$L_{1,2}^{\{1\}} = bL_{2,2}^{\{1\}} = b(L_{2,2}^{\{1,2\}})^{*} = b\{b, abb\}^{*}.$$

Stąd otrzymujemy równość

$$\rho_2 = b(b + abb)^{*}.$$

Wyliczenie wyrażenia ρ_1 pozostawiam zainteresowanym Czytelnikom.

5 Kolejne rozwiązanie

Wprowadźmy teraz drobne pojęcie. Blokiem ab w słowie w będziemy nazywać dowolne podsłowo słowa w, które zaczyna się literą a i kończy się literą b, w którym wszystkie litery a znajdują się przed literami b, spełniające warunek maksymalności, a więc takie podsłowo, którego nie można powiększyć zachowując wszystkie wymienione tutaj własności. Na przykład, słowo $baabbbab$ zawiera dokładnie dwa bloki ab: pierwszy, złożony z dwóch liter a i trzech liter b, i drugi, złożony z dwóch ostatnich liter ab. Podsłowo ab rozważanego słowa, złożone z trzeciej i czwartej litery, nie jest blokiem ab gdyż nie jest maksymalnym podsłowem o własnościach wymaganych przez definicję bloku.

Lemat 5.1 *Przypuśćmy, że słowo w zaczyna się literą a i kończy się literą b. Słowo w zawiera wzorzec $baba$ wtedy i tylko wtedy, gdy zawiera zawiera dwuliterowy blok ab. Wobec tego słowo w nie zawiera wzorca $baba$ wtedy i tylko wtedy, gdy każdy zawarty w nim blok ab ma przynajmniej trzy litery.* □

Lemat ten można i trzeba uogólnić na przypadek dowolnych słów, w których pewna litera b znajduje się przed literą a. Wtedy będzie można go użyć do znalezienia wyrażenia regularnego reprezentującego język złożonych ze słów nie zawierających wzorca $baba$, ale zawierających literę b przed a. Wyrażeniem przypominającym to szukane, oczywiście błędnym, może być

$$a^*b^*b(a^*aabb^*)^*aa^*b^*.$$

Teraz, do znalezienia pełnego rozwiązania, brakuje już tylko drobiazgu.