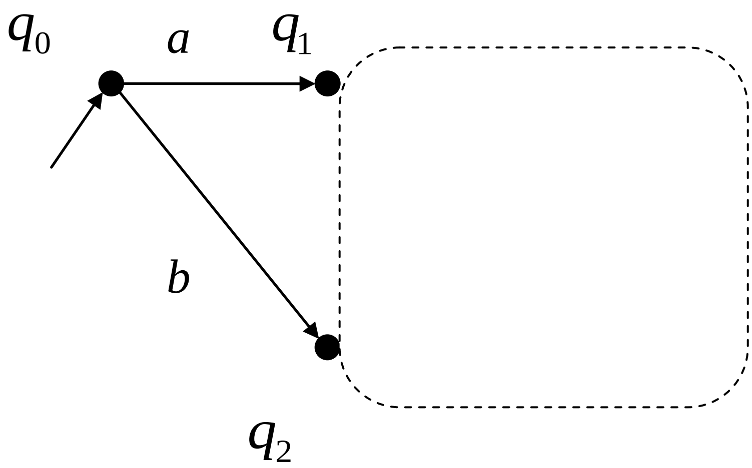
§8. Поиск регулярного выражения по автомату. Теорема Ардена. Решение системы уравнений с языками.

Задача: найти регулярное выражение для языка L, допускаемого ДКА (возможно неполным).

Замечание: в доказательстве теоремы Клини нет алгоритма поиска регулярного языка по автомату.

Пример. $\Sigma = \{a, b\}$



Обозначим S_0 – язык слов автомата, где q_0 – начальное состояние;
S_1 – язык слов автомата, где q_1 – начальное состояние;
S_2 – язык слов автомата, где q_2 – начальное состояние.

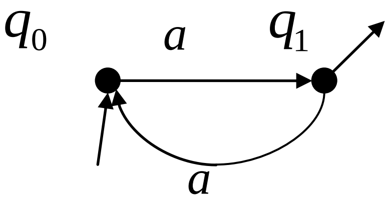
Тогда $S_0 = a \cdot S_1 + b \cdot S_2$.

Аналогично, S_1 и S_2 можно выразить через $\{S_i\}$, где
S_i – язык слов автомата, где q_i – начальное состояние.

Для заключительных состояний можно получить выражение, не содержащее S_i в правой части.

Проблемы возникают, когда для выражения S_i получаем в правой части тоже S_i .

Пример.



$$\begin{cases} S_0 = a \cdot S_1 \\ S_1 = a \cdot S_0 + \varepsilon \end{cases}$$

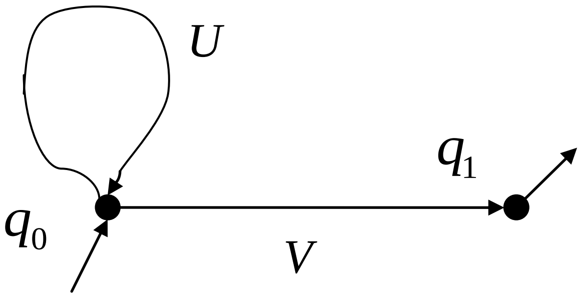
$$\Rightarrow \; S_0 = a \cdot (a \cdot S_0 + \varepsilon).$$

Воспользуемся догадкой: $S_0 = a \cdot \left(a^2\right)^* = \left(a^2\right)^* \cdot a.$

Теорема (Ардена).
Для любых регулярных языков U, X, V: $\varepsilon \notin U$, уравнение $X = U \cdot X + V$ имеет единственное решение $X = U^* \cdot V$.

Замечание: уравнению $X = X + V$ удовлетворяет любой язык X, для которого $V \subseteq X$.

Иллюстрация для запоминания теоремы Ардена:



Доказательство теоремы Ардена:

1) Подставим $X = U^* \cdot V$ в уравнение:

$U^* \cdot V = U \cdot U^* \cdot V + V$,

$U^* \cdot V = (U \cdot U^* + \varepsilon) \cdot V$ – верное равенство.

2) Пусть X' удовлетворяет уравнению $X = U \cdot X + V$.

Тогда $X' = U \cdot (U \cdot X' + V) + V = U^2 \cdot X' + (U + \varepsilon) \cdot V$.

$X' = U^2 \cdot (U \cdot X' + V) + V = U^3 \cdot X' + (U^2 + U + \varepsilon) \cdot V$.

$$\forall n \in \mathrm{N}$$

$$X' = U^{n+1} \cdot X' + (U^{n} + \ldots + U + \varepsilon) \cdot V \; .$$

$$\forall w \in U^{*} \cdot V \rightarrow \; w \in X' \; .$$

Осталось показать, что $X' \subseteq U^* \cdot V$.

(от противного) Предположим, $\exists w \in X' \colon\ w \notin U^* \cdot V$, w – кратчайшее.
$w \in UX',\ \ w \notin V$.
Тогда $w = u \cdot w'$, где $u \neq \varepsilon$, $u \in U,\ \ w' \in X'$.
$|w'| < |w| \ \Rightarrow w' \in U^* \cdot V$.

Тогда $w \in U^n \cdot V \subseteq U^* \cdot V$ – противоречие.

Следовательно, $X = U^* \cdot V$ – единственное решение. Теорема доказана.

Алгоритм поиска регулярного выражения:

1) Составим систему уравнений для S_i:

$$S_i = \sum_{a_j \in \Sigma} a_j \cdot S_m \text{, если } q_i \notin Q_F \text{;}$$

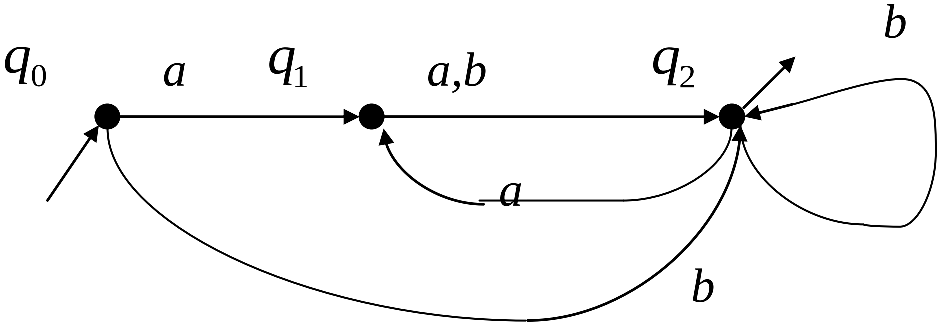
$$S_i = \varepsilon + \sum_{a_j \in \Sigma} a_j \cdot S_m \text{, если } q_i \in Q_F \text{;}$$

где $\varphi(q_i, a_j) = q_m$.

2) Удаляем последовательно S_i из системы.

3) S_0 – ответ.

Пример. Автомат для продажи кофе, где монета 5 соответствует символу *a*, монета 10 соответствует символу *b*.



$$\begin{cases} S_0 = a\cdot S_1 + b\cdot S_2 \\ S_1 = (a+b)\cdot S_2 \\ S_2 = \varepsilon + a\cdot S_1 + b\cdot S_2 \end{cases}$$

$$\begin{cases} S_0 = a \cdot S_1 + b \cdot S_2 \\ S_1 = (a+b) \cdot S_2 \\ S_2 = \varepsilon + a \cdot S_1 + b \cdot S_2 \end{cases}$$

Подставим S_1 в первое и третье уравнения:

$$\begin{cases} S_0 = a \cdot (a+b) \cdot S_2 + b \cdot S_2 \\ S_2 = \varepsilon + a \cdot (a+b) \cdot S_2 + b \cdot S_2 \end{cases}$$

Упростим уравнения:

$$\begin{cases} S_0 = \left[a \cdot (a+b) + b\right] \cdot S_2 \\ S_2 = \varepsilon + \left[a \cdot (a+b) + b\right] \cdot S_2 \end{cases}$$

$$\begin{cases} S_0 = \left[a \cdot (a+b) + b\right] \cdot S_2 \\ S_2 = \varepsilon + \left[a \cdot (a+b) + b\right] \cdot S_2 \end{cases}$$

Выразим S_2 из второго уравнения: $S_2 = (a(a+b)+b)^* \cdot \varepsilon$.

Упростим $S_2 = (a(a+b)+b)^*$

Подставим в первое уравнение:

$S_0 = (a(a+b)+b) \cdot (a(a+b)+b)^*$.

Запишем в более короткой форме:

$S_0 = (a(a+b)+b)^n$, где n $\in$ **N** – ответ.

В оригинальной статье Ардена использовалась альтернативная формулировка теоремы при альтернативных обозначениях.

Пусть T_i – язык слов автомата, где q_i – единственное заключительное состояние. Тогда уравнение с языками имеет вид:

$T_i = \sum_{a_j \in \Sigma} T_m \cdot a_j$, если q_i не является начальным;

$T_i = \varepsilon + \sum_{a_j \in \Sigma} T_m \cdot a_j$, если q_i – начальное состояние;

где $\varphi(q_m, a_j) = q_i$.

Альтернативная теорема Ардена.
Для любых регулярных языков U, X, V: $\varepsilon \notin U$, уравнение $X = X \cdot U + V$ имеет единственное решение

$$X = V \cdot U^*.$$

После решения системы уравнений нужно найти T_i для каждого состояния из Q_F. Ответом будет их сумма.

§9. Лемма о накачке. Пример нерегулярного языка.

Лемма о разрастании, pumping lemma.

Лемма о накачке является важным теоретическим результатом, позволяющим проверить, является ли данный язык регулярным.

Лемма применяется для проверки бесконечных языков, т.к. любой конечный язык – регулярный.

Лемма (о накачке). Для бесконечного регулярного языка L над алфавитом $\sum$ выполняется:

$$\exists n \in \mathbf{N}\text{:}\ \forall \alpha \in L :\ |\alpha| \geq n\ \exists u, v, w \in \Sigma^{*} : \alpha = uvw,\ \ |uv| \leq n,\ \ |v| \geq 1,$$

$$\forall i \in \mathbf{N} \cup \{0\}\ \rightarrow\ uv^{i}w \in L.$$

Доказательство леммы о накачке:

Пусть язык L допускается ДКА (Q, Σ, φ, q_0, Q_F).

Пусть $|Q| = n$. Рассмотрим слово $\alpha \in$ L: $|\alpha| \geq n$.

$q_0, q_1, \cdots q_n, \cdots, q_{|\alpha|}$ – последовательность состояний, по которым переходит автомат при просмотре букв слова α.

В соответствии с принципом Дирихле существуют индексы j и k:

$0 \leq j < k \leq n$, $q_j = q_k$.

Выберем в слове α подслова u, v. w: $|u| = j$, $|v| = k - j$, $\alpha = uvw$.

Тогда слово uw допускается этим же ДКА.

$q_0, q_1, \cdots q_j, q_{k+1}, \cdots, q_{|\alpha|}$ – последовательность состояний, по которым переходит автомат при просмотре букв слова uw.

И для любого натурального числа i слово $uv^i w$ допускается этим же ДКА.

$q_0, q_1, \dots q_j, (q_{j+1}, \dots, q_k)^i, q_{k+1}, \dots, q_{|\alpha|}$ – последовательность состояний, по которым переходит автомат при просмотре букв такого слова. Лемма доказана.

Замечание: лемма о накачке дает необходимое условие регулярности языка, но не достаточное.

Для ее применения часто используют обратное утверждение:
Если $\forall n \in \mathbf{N}$:

$$\exists \alpha \in L : \ |\alpha| \geq n \ \forall u, v, w \in \Sigma^* : \alpha = uvw, \ |uv| \leq n, \ |v| \geq 1,$$

$$\exists i \in \mathbf{N} \cup \{0\} : \quad uv^i w \notin L,$$

то язык L не регулярный.

Пример нерегулярного языка.

Пусть $\Sigma = \{0, 1\}$, $L = \{\, p\, c\, R(p)\}$, где $p \in \Sigma^*$, $c \in \Sigma$,

$R(p)$ – слово, полученное из p записью букв в обратном порядке.

Для любого n существует слово $\alpha = \underbrace{0...0}_{n}1\underbrace{0...0}_{n} \in L$, длина которого равна $2n+1$.

Для любого деления слова α на части uvw, чтобы $|uv| \le n,\ |v| \ge 1$ выполняется: $v = \underbrace{0...0}_{k},$. Тогда $uv^2w = \underbrace{0...0}_{n+k}1\underbrace{0...0}_{n} \notin L$.

Следовательно, язык L не регулярный.