1 Регулярные языки

Определение 1.1. *Возьмем некоторое конечное множество символов A, назовем его алфавитом, а его элементы — буквами.*

Определение 1.2. *Словом в данном алфавите называется конечная цепочка букв этого алфавита.*

Буквы будем обозначать $a, a_1, a_2, \ldots, b, b_1, \ldots$, а слова – $\alpha, \beta, \ldots$ Будем обозначать через $\alpha[i]$ – i-ю букву слова α. Таким образом, $\alpha = \alpha[1]\alpha[2]\ldots\alpha[n]$.

Определение 1.3. *Длиной слова α называется число букв в данном слове $|\alpha| = n$.*

Например, $|abbbc| = 5$.

Определение 1.4. *Введем также пустое слово Λ как слово нулевой длины $|\Lambda| = 0$.*

Определение 1.5. *Слово β называется подсловом слова α, если найдутся слова α_1 и α_2, необязательно непустые, что $\alpha = \alpha_1\beta\alpha_2$.*

Например, подсловами слова abc является само abc, а также a,b,c,ab и bc. Множество всех возможных слов в алфавите A обозначим A^*.

Определение 1.6. *Языком в данном алфавите A называется любое подмножество L множества всех слов A^*, $L \subseteq A^*$.*

Пример.
$A = \{a, b, c\}$.
$L = \{\Lambda, aa, abc, cb, bc\}$.
Множество A^* – все слова, которые можно составить из букв a,b,c: $\Lambda, a, b, c, aa, ab, ac, ba, \ldots$.

2 Операции над языками

Рассмотрим произвольный алфавит A и всевозможные языки в нем.
Определим следующие операции.

Определение 2.1. *Объединением языков L_1 и L_2 называется множество слов, входящих хотя бы в один из этих языков*

$$L = L_1 \cup L_2 = \{\alpha | \alpha \in L_1 \vee \alpha \in L_2\}.$$

Определение 2.2. *Конкатенацией языков L_1 и L_2 называется множество слов вида*

$$L_1 \cdot L_2 = \{\alpha\beta | \alpha \in L_1, \beta \in L_2\}.$$

Таким образом, это слова, получающиеся приписыванием к каждому слову из L_1 слова из L_2.
Конкатенация слов α и β есть слово $\alpha\beta$.
Например, пусть $L_1 = \{a, ab, b\}$, $L_2 = \{b, ca\}$.
Тогда

$$L_1 \cup L_2 = \{a, ab, b, ca\}$$

и

$$L_1 \cdot L_2 = \{ab, abb, bb, aca, abca, bca\}.$$

В частности, $\underbrace{L \dots L}_{k}$ – конкатенация языка k раз обозначается как L^k
и есть $\{\alpha_1 \dots \alpha_k | \alpha_i \in L, i = 1 \dots k\}$.
Например, $L = \{a, bb\}$, $L^2 = \{aa, abb, bba, bbbb\}$.
Рассмотрим произвольный язык L и пустое слово Λ.
По определению $\Lambda \cdot L = L$ и $L \cdot \Lambda = L$.
В качестве языка можно рассматривать и пустое множество слов.
Выполнено:
$L \cdot \varnothing = \varnothing \cdot L = \varnothing$.
$L \vee \varnothing = L$.

Определение 2.3. *Итерацией языка L называется язык вида*

$$L^* = \Lambda \cup L \cup L^2 \cup \dots \cup L^i \cup \dots .$$

Например, в алфавите $A = \{a, b\}$ итерация языка $L = a^2, ab$ будет

$$L^* = \{\Lambda, a^2, ab, a^4, abab, a^3b, ab^2a, a^6, a^5b, aba^4, \dots\}.$$

Для $L = \{a^2\}$ итерация такова

$$L^* = \{\Lambda, a^2, a^4, a^6, \dots\}.$$

Итерация пустого множества есть пустое слово $\varnothing^* = \Lambda$.
Множество всех слов в алфавите $A = \{a_1, \dots, a_r\}$ получается итерацией объединения его букв $A^* = (a_1 \cup \dots \cup a_r)^*$.

Определение 2.4. *Языки $\{\Lambda\}$, $\{a | a \in A\}$, а также пустое множество слов $\varnothing$, называются простейшими языками.*

Определение 2.5. *Язык называется регулярным, если его можно получить из простейших языков с помощью этих трех операций за конечное число шагов.*

Определение 2.6. *Символьное выражение, задающее регулярный язык, называется регулярным выражением.*

Примеры

1. **Задача**

   Составить регулярное выражение для языка в алфавите $\{a, b, c\}$, состоящее из всех слов, начинающихся на ab, но не заканчивающихся на c.

   Решение:

   Как уже было сказано, множество всех слов в алфавите $A = \{a, b, c\}$ есть $A^* = (a \cup b \cup c)^*$.

Все слова, начинающиеся на ab – конкатенация ab с множеством всех слов.

Выражение для такого языка есть $ab(a \cup b \cup c)^*$.

Слово не заканчивается на букву c, значит, оно заканчивается на a или на b.

Поэтому регулярное выражение для данного языка имеет вид $ab(a \cup b \cup c)^*(a \cup b)$.

2. **Задача**

   Составить регулярное выражение для языка в алфавите $\{a, b, c\}$ из всех слов, где буква b встречается только в виде массива b^n, где n – четное число.

   Решение:

   Сначала зададим массив b^n. Это $(bb)^*$.

   Слова языка – всевозможные последовательности букв a,c и таких массивов.

   Искомое регулярное выражение – $(a \cup (bb)^* \cup c)^*$.

3. **Задача** Дан язык L в алфавите $A = \{a, b, c\}$. Записать регулярное выражение для языка, у всех слов которого, на всех нечетных местах находится буква a.

   Решение:

   Рассмотрим слово α длины $|\alpha| = n$.

   Возможны два варианта:

   - n – чётное.

     По условию на нечётных местах находится буква a, а на чётных местах может быть любая буква алфавита, в том числе и a.

1	2	3	4	5	6	...
a	$a \cup b \cup c$	a	$a \cup b \cup c$	a	$a \cup b \cup c$	...

     Рассмотрим конструкцию из двух букв: первую букву запишем как a, вторая буква может быть любой: a, b или c, эту конструкцию можем записать с помощью операций объедиения и конкатенации

     $$L_0 = a(a \cup b \cup c).$$

     Все возможные слова чётной длины, удовлетворяющие условию получим с помощью итерации

     $$L_1 = L_0^* = (a(a \cup b \cup c))^* = \Lambda \cup a(a \cup b \cup c) \cup (a(a \cup b \cup c))^2 \cup \dots$$

   - n – нечётное.

     Для того чтобы учесть вариант, когда длина слова нечётная, добавим ещё один символ a.

     $$L_2 = L_1 \cdot a.$$

     Случай, когда формируется слово длины 1 получается при конкатенации пустого слова Λ и a.

Таким образом, интересующий нас язык образует объединение слов чётной и нечетной длины

$$L = L_1 \cup L_2 = (a(a \cup b \cup c))^* \cup (a(a \cup b \cup c))^* a = (a(a \cup b \cup c))^* (\Lambda \cup a)$$

3 Источники

Определение 3.1. *Пусть зафиксирован некоторый алфавит A. Возьмем ориентированный псевдограф, некоторым ребрам которого приписаны буквы из алфавита A. Выделим некоторое множество вершин, называемых начальными и множество вершин, называемых заключительными. Такая конструкция называется источником.*

Определение 3.2. *Ребра без букв назовем пустыми.*

Начальные вершины обозначаются $*$, а заключительные $\bullet$.

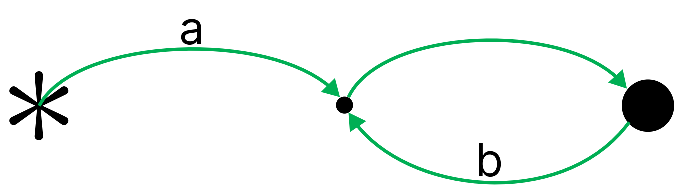

Рис. 1: Пример источника

Рассмотрим путь $e_1, \ldots, e_k$ в источнике. Выпишем последовательно буквы, приписанные рёбрам $e_1, \ldots, e_k$: $a_1, \ldots, a_k$. Получившееся слово назовём словом, порожденным данным путем. Если все рёбра пути пустые, то такой путь порождает пустое слово.
Каждому источнику ставится в соответствие язык $L \subseteq A^*$ следующим образом. Для каждого пути из некоторой начальной вершины в некоторую заключительную выписывается порожденное им слово. Все такие слова, и только они составляют язык L. Говорят, что источник порождает язык L.
Чтобы проверить, что данный источник порождает именно этот язык, нужно рассмотреть все пути, ведущие из начальной вершины в заключительную.

Определение 3.3. *Источники называются эквивалентными, если они порождают один и тот же язык.*

Определение 3.4. *Источник называется двухполюсником, если в нем ровно одна начальная вершина q_0 и ровно одна заключительная q_f такие, $q_0 \neq q_f$, $\deg^-(q_0) = 0$ и $\deg^+(q_f) = 0$.*

Утверждение 3.5. *Для любого источника существует эквивалентный ему двухполюсник.*

Доказательство. В данном источнике все начальные и заключительные вершины сделаем обыкновенными и введем дополнительные вершины q_0 и q_f. Из q_0 проведем пустые ребра в бывшие начальные, а из бывших заключительных проведем пустые ребра в q_f. Получившийся источник – двухполюсник, эквивалентный данному. □

Лемма 3.6. *Пусть вершины источника пронумерованы, а R_{ij}^k – множество всех слов, порожденных путями в данном источнике из вершины с номером i в вершину с номером j, ранг которых не превосходит k.*
Тогда справедливы следующие утверждения:

1. $R_{ij}^k = R_{ij}^{k-1} \cup R_{ik}^{k-1}(R_{kk}^{k-1})^* R_{kj}^{k-1}$.

2. $R_{ik}^k = R_{ik}^{k-1}(R_{kk}^{k-1})^*$.

3. $R_{kj}^k = (R_{kk}^{k-1})^* R_{kj}^{k-1}$.

4. $R_{kk}^k = (R_{kk}^{k-1})^*$.

Эта лемма позволяет от R_{ij}^k перейти к простейшим языкам R_{ij}^0.

Теорема 3.7 (Теорема Клини для источников)**.** *Каждый язык, порождаемый источником, является регулярным.*

Доказательство. Составим регулярное выражение для языка, порожденного произвольным источником.
Итак, рассмотрим источник **И** с n вершинами. Некоторым образом перенумеруем его вершины. Множество начальных вершин обозначим I, а множество заключительных – F.
Очевидно, что вырабатываемое им множество слов есть

$$\bigcup_{l \in I, k \in F} R_{lk}^n.$$

Для каждого R_{lk}^n применяем лемму до тех пор, пока в ней не будут участвовать лишь R_{ij}^0, то есть слова, соответствующие множествам путей из вершины i в вершину j, не заходящих ни в какую другую вершину.
Можно выписать конкретные выражения для каждого R_{ij}^0 следующим образом.

- $R_{ij}^0 = \varnothing$, если нет ребер, ведущих из вершины i в вершину j, причем $i \neq j$.

- $R_{ij}^0 = a_1 \cup \dots \cup a_k$, если из i в j ведут ребра с буквами $a_1, \dots, a_k$. Если от i к j ведет еще и пустое ребро, то в объединение добавляется пустое слово Λ.

- $R_{ij}^0 = \Lambda$, если есть только пустое ребро.

- Множество R_{ii}^0 также всегда содержит пустое слово Λ.

Ясно, что языки R_{ij}^0 регулярны. Видно, что все языки R_{ij}^k получаются из них с помощью операций объединения, конкатенации и итерации. Следовательно, все языки R_{ij}^k регулярны, поэтому регулярен и язык $\bigcup\limits_{l \in I, k \in F} R_{lk}^n$.

□

Пример

Пусть L означает язык, порождённый данным источником. Выразить L при помощи регулярного выражения.

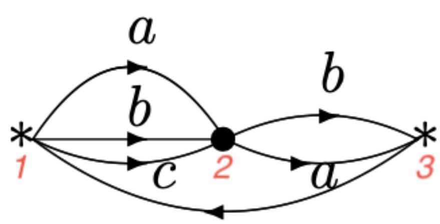

Рис. 2: Источник

Решение

Основные формулы, позволяющие снижать ранг пути, имеют вид

$$R_{ij}^{k} = R_{ij}^{k-1} \cup R_{ik}^{k-1}(R_{kk}^{k-1})^{*}R_{kj}^{k-1}, \tag{1}$$

$$R_{kj}^{k} = (R_{kk}^{k-1})^{*}R_{kj}^{k-1}, \tag{2}$$

$$R_{ik}^{k} = R_{ik}^{k-1}(R_{kk}^{k-1})^{*}, \tag{3}$$

$$R_{kk}^{k} = (R_{kk}^{k-1})^{*}. \tag{4}$$

Интересующий нас язык получается при прохождении из всех начальных вершин во все конечные

$$L = R_{12}^{3} \cup R_{32}^{3}.$$

Применим формулу (1) к выражению R_{12}^{3}

$$R_{12}^{3} \overset{(1)}{=} R_{12}^{2} \cup R_{13}^{2}(R_{33}^{2})^{*}R_{32}^{2}.$$

Мы получили выражение через пути ранга 2, рассмотрим их по отдельности:

1. $R_{12}^{2} \overset{(3)}{=} R_{12}^{1}(R_{22}^{1})^{*}$.

   Распишем каждый элемент в этой формуле через простейшие языки:

   - $R_{12}^{1} \overset{(2)}{=} (R_{11}^{0})^{*}R_{12}^{0} = \Lambda^{*}(a \cup b \cup c) = a \cup b \cup c.$
   - $R_{22}^{1} \overset{(1)}{=} R_{22}^{0} \cup R_{21}^{0}(R_{11}^{0})^{*}R_{12}^{0} = \Lambda \cup \emptyset = \Lambda$

   Тогда получим $R_{12}^{2} = a \cup b \cup c$.

2. $R_{13}^{2} \overset{(1)}{=} R_{13}^{1} \cup R_{12}^{1}(R_{22}^{1})^{*}R_{23}^{1}$.

   Выражения для R_{12}^{1} и R_{22}^{1} уже получены, распишем оставшиеся элементы через простейшие языки:

   - $R_{13}^{1} \overset{(2)}{=} (R_{11}^{0})^{*}R_{13}^{0} = \Lambda^{*}\emptyset = \emptyset.$

- $R^1_{23} \overset{(1)}{=} R^0_{23} \cup R^0_{21}(R^0_{11})^* R^0_{13} = a \cup b \cup \emptyset = a \cup b.$

Тогда получим $R^2_{13} = (a \cup b \cup c)(a \cup b)$.

3. $R^2_{33} \overset{(1)}{=} R^1_{33} \cup R^1_{32}(R^1_{22})^* R^1_{23}$.

   Распишем R^1_{33} и R^1_{32}:

   - $R^1_{33} \overset{(1)}{=} R^0_{33} \cup R^0_{31}(R^0_{11})^* R^0_{13} = \Lambda \cup \emptyset = \Lambda.$
   - $R^1_{32} \overset{(1)}{=} R^0_{32} \cup R^0_{31}(R^0_{11})^* R^0_{12} = a \cup b \cup c.$

   Тогда получим $R^2_{33} = (a \cup b \cup c)(a \cup b)$.

4. $R^2_{32} \overset{(3)}{=} R^1_{32}(R^1_{22})^*$.

   Все необходимые выражения были получены ранее, поэтому $R^2_{32} = a \cup b \cup c$.

Подставим найденные выражения в формулу для R^3_{12} и упростим, после чего получим

$$R^3_{12} = ((a \cup b \cup c)(a \cup b))^*(a \cup b \cup c).$$

Теперь применим формулу (2) к R^3_{32}.

$$R^3_{32} \overset{(2)}{=} (R^2_{33})^* R^2_{32}$$

Все необходимые выражения были получены ранее, поэтому

$$R^3_{32} = ((a \cup b \cup c)(a \cup b))^* \,(a \cup b \cup c).$$

Теперь можно записать выражение для L, после упрощения получим

$$L = ((a \cup b \cup c)(a \cup b))^*(a \cup b \cup c).$$

4 Порождающие грамматики

Определение 4.1. *Порождающей грамматикой называется четвёрка*

$$G = (T, N, I, P),$$

где T – терминальный алфавит,
N – нетерминальный алфавит, причём $T \cap N = \varnothing$,
I – выделенный символ нетерминального алфавита (аксиома),
P – конечное множество правил вывода (продукция), причём $P \subseteq (T \cup N)^+ \times (T \cup N)^*$.

Пары $(\alpha, \beta) \in P$ называются правилами вывода, просто правилами или продукциями и записывают как $\alpha \to \beta$.
Для обозначения n правил с одинаковыми левыми частями $\alpha \to \beta_1, \ldots, \alpha \to \beta_n$ часто используют сокращённую запись $\alpha \to \beta_1 | \ldots | \beta_n$.
Пример.
Пусть даны множества $N = \{I\}$, $T = \{(,)\}$,

$P = \{I \to (I), I \to II, I \to \Lambda\}$.
Тогда (T, N, I, P) является порождающей грамматикой, задающей правильную скобочную последовательность.
Вывод строки «(()())»:
$I \to (I) \to (II) \to ((I)I) \to ((I)(I)) \to (()(I)) \to (()())$.

Иерархия Хомского

1. Грамматики типа 0

   К этому классу относятся все формальные грамматики

2. Грамматики типа 1 (контекстно-зависимые)

   Правила вывода имеют вид:

   $$\alpha A\beta \to \alpha\gamma\beta,$$

   где $\alpha, \beta \in (T \cup N)^*, A \in N, \gamma \in (T \cup N)^+$.

3. Грамматики типа 2 (контекстно-свободные)

   Правила вывода имеют вид:

   $$A \to \beta,$$

   где $A \in N$, $\beta \in (T \cup N)^*$.

4. Грамматики типа 3 (регулярные)

   Правила вывода имеют вид:

   $$A \to \gamma B \text{ или } A \to \gamma,$$

   где $A, B \in N$, $\gamma \in T$.