**Свойства регулярных выражений и языков**

1. **Формальные языки**

Пусть **Σ** – не пустое конечное множество, называется алфавитом.

|**Σ**| = n > 0 – мощность алфавита.

Произвольный элемент **а ϶ Σ** называется символом. На множестве **Σ** определен линейный порядок. Каждый символ **а ϶ Σ** имеет уникальный числовой код **k(a)** на отрезке [0-(n-1)].

Если **a,b ϶ Σ** и **a ≠ b**, тогда **a < b** , если **k(a) < k(b)**.

Словом (абстрактным) над алфавитом Σ подразумевается конечная последовательность символов, возможно повторяющихся в различных позициях: w = w0 w1… wm-1 , где длина слова |w| = m > 0, wi – произвольный символ, i ϶ [0-( m-1)].

Пустое слово имеет длину 0 и обозначается символом λ ∌ Σ. Множество всех слов, включая пустое слово, обозначается Σ\* .

На множестве Σ\* определена двуместная операция сцепления (конкатенации) двух слов:

если w = w0 w1… wm-1 и v = v0 v1… vk-1 , тогда сцепление

wv = w0 w1… wm-1 v0 v1… vk-1 – слово длины |wv| = m + k, m > 0 , k > 0.

Пустое слово λ выполняет роль еденицы, поскольку верны следующие тождества λw = wλ = w , где w – любое слово языка Σ\*.

Множество Σ\*, замкнутое относительно операции сцепления, и, с пустым словом в качестве еденицы, является моноидом.

Любое подмножество слов L ⊆ Σ\* называется формальным языком.

Если L – конкретный (определенный) язык, его элементы называются предложениями.

Множество Σ\* называется универсальным языком в алфавите Σ.

Над языками, как множествами, определены теоретико-множественные операции, производящие языки:

если L и M - языки, тогда L ⋃ M – объединение, L ⋂ M – пересечение,

L \ M – разность – формальные языки.

Пустой язык обозначается символом пустого множества ⊘.

Определена также операция сцепления двух языков:

L M = {wv| w ϶ L, v ϶ M}.

Пустой язык ⊘ и язык Λ = {λ} играют роль аддитивной и мультипликативной еденицы в семействе всех формальных языков, соответственно:

⊘ ⋃ L = L ⋃ ⊘ = L ; ΛL = LΛ=L.

Для операций объединения и сцепления языков имеют место ассоциативные законы, соответственно:

(L ⋃ M) ⋃ S = L ⋃ ( M ⋃ S) = L ⋃ M ⋃ S,

(LM)S = L(MS) = LMS, где L, M, S – произвольные языки в алфавите Σ.

Для операций объединения и сцепления языков имеют место также два дистрибутивных закона:

(L ⋃ M)S = (LS) ⋃ (MS) = LS ⋃ MS,

L(M ⋃ S) = (LM) ⋃ (LS) = LM ⋃ LS.

По определению L0 = Λ = {λ} для всех языков L , включая пустой язык.

Степень языка Li при i >1 определяется по формуле:

Li = Li-1 L.

Итерация языков определяется унарной постфиксной операцией звездой Клини:

L\* = L0 ⋃ L1 ⋃ … ⋃ Li ⋃ …

или

L\* = .

Позитивная итерация определяется по формуле

L+ = .

Следующие тождества верны для всех языков:

L ⋃ L = L,

Λ ⋃ L+  = L\*,

(Λ ⋃ L)+ = L\*,

(Λ ⋃ L)\*  = L\*,

L\*\* = L\*.

2. **Тождества для регулярных выражений**

Пусть e, e1, e2, e3, e4 – произвольные регулярные выражения над алфавитом Σ.

Верны следующие тождества:

⊘|e = e|⊘,

λe = eλ, λλ = λ;

дистрибутивные законы:

(e1|e2)e3 = (e1e3)|(e2e3) = e1e3|e2e3,

e1(e2|e3) = (e1e2)|(e1e3) = e1e2|e1e3;

ассоциативные законы:

(e1|e2)|e3 = e1|(e2|e3),

(e1e2)e3 = e1(e2e3);

e|e = e,

λ|e+ = e\*,

(λ|e)+ = e\*,

(λ|e)\* = e\*,

e\*\* = e\*.

Все выше перечисленные тождества справедливы, поскольку они определяют

регулярные языки, как подсемейство формальных языков и, поэтому проверяются непосредственно.

3. Специальные расширения регулярных выражений

Регулярное выражение вида

а1| а2| … |аm ,

где аi ϶ Σ, 1≤ i ≤ m,

называется элементарным аддитивным выражением.

Это выражение определяет элементарный не пустой язык

М = {а1, а2, … ,аm} ⊆ Σ, состоящий из одно-символьных слов.

По общепринятому соглашению одно-символьные слова обозначаются символами алфавита.

Язык М называется элементарным языком, является некоторым не пустым подмножеством алфавита Σ.

Чтобы компактно определить произвольный элементарный язык М ⊆ Σ, используются метасимволы [,] – квадратные скобки.

1. Язык М = {c| a ≤ c ≤ b} определяется выражением [a-b], которое называется отрезком алфавита.

Тождества иллюстрируют применение [a-b]:

0|1|2|3|4|5|6|7|8|9 = [0-9],

A|B|C|D|F = [A-F],

x|y|z = [x-z],

0|1 = [0-1];

a = [a] = [a-a], если a = b.

Объединение отрезков:

[0-9]|[A-F] = [0-9A-F] = [A-F]| [0-9],

[0-9]|[A-Z]|[a-z]|\_ = [0-9A-Za-z\_] = [\_A-Za-z0-9]

x|X = [xX] = [Xx],

0|1|2|3|4|5|6|7|8|9 = [0-9] = [012-89] =[ 0123456789].

Ввиду коммутативности операции объединения «|» внутри квадратных скобок отрезки можно записывать в любом порядке.

Метасимволы (, ), +, \*, [ , ] в регулярных выражениях, представляющие объектные символы алфавита Σ, записываются с помощью искейпт-последовательностей, например, \(, \\*, \].

Любые символы алфавита Σ можно записывать в регулярных выражениях с помощью числовых искейпт-последовательностей:

в системе счисления с основанием 8 -- в форме

\octal 2-3 digit octal character code, например,

символ плюс + записывается либо \+ либо \053,

так как его порядковый номер (числовой код) в ASCII-таблице – 053;

символ буква F записывается либо F либо \106.

В системе счисления с основанием 16 символ плюс + имеет запись \x2B,

а символ буква F - \x46.

Объединение отрезков можно выразить различными эквивалентными способами, например,

[\_A-Za-z0-9] = [\5F\x41-\x5Aa-z0-9].

Пусть M – элементарный язык определяется выражением в форме [R], где R обозначает запись отрезков алфавита определяющих множество M.

Тогда разность Σ \ M, как элементарный язык, определяется выражением в форме [^R]. Например, [^0-9] - элементарный язык одно-символьных слов образован не из цифр. Говорят, что алфавит имеет разбиение на два класса

С1 = [R] и С2 = [^R].

4. Разбиение алфавита

Пусть M1, M2, … , Mk , k ≥ 2, - семейство попарно не пересекающихся элементарных языков таких, что Σ = M1 ⋃ M2 ⋃… ⋃ Mk и

Mk = Σ \ (M1 ⋃ M2 ⋃… ⋃ Mk-1).

Пусть [Mi] – специальное регулярное выражение для языка Mi , 1≤ i < k,

тогда только язык Mk определяется выражением в форме [^R], где

R = M1 ⋃ M2 ⋃… ⋃ Mk-1.

Ниже в качестве примера приводятся два разбиения Π1 и Π2

алфавита Σ = ASCII:

Π1 = {[\_A-Za-z], [0-9], [^\_A-Za-z0-9]}, k =3.

Π2 **=** {[0], [1-7], [lL], [uU], [^01-7 lL uU]}, k = 5.

5. **Пересечение разбиений**

На множестве всех разбиений алфавита Σ определена операция пересечения

двух разбиений.

Пусть П1 = {M1, M2 , … , Mk}, k ≥ 1, и П2 = {N1, N2 , … , Nm}, m ≥ 1,

пересечение двух разбиений П1 и П2 определяется следующим образом:

П12 = П1 ⋂ П2 = { Mi ⋂ Nj ≠ ⊘| Mi ∊ П1, 1 ≤ i ≤ k, и Nj ∊ П2 , 1 ≤ j ≤ m}.

Пусть П12 = {R1, R2, … ,Rp}, p ≥ 1.

Каждый класс Mi ∊ П1 однозначно представим как объединение некоторых классов разбиения П12. Аналогично, каждый класс Nj ∊ П2 однозначно представим как объединение некоторых классов разбиения П12.

6. **Вычисление пересечения разбиений**

Рассмотрим представление классов разбиений с помощью элементарных аддитивных выражений в форме [M] и на примере двух разбиений алфавита

Σ = ASCII

Π1 = {[\_A-Za-z], [0-9], [^\_A-Za-z0-9]}, k =3,

Π2 **=** {[0], [1-7], [lL], [uU], [^01-7 lL uU]}, k = 5,

вычислим П12 = П1 ⋂ П2.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  | N1 | N2 | N3 | N4 | N5 |
|  | П1 ⋂ П2 | [0] | [1-7] | [lL] | [uU] | [^01-7 lL uU] |
| M1 | [\_A-Za-z] |  |  | [lL] | [uU] | [\_A-KM-TV\_Za-km-tv-z] |
| M2 | [0-9] | [0] | [1-7] |  |  | [89] |
| M3 | [^\_A-Za-z0-9] |  |  |  |  | [^\_A-Za-z0-9] |

Пустые клетки таблицы соответствуют пустому множеству, если Mi ⋂Nj = ⊘.

П1 ⋂ П2 = {[0], [1-7], [lL], [uU], [\_A-KM-TV\_Za-km-tv-z], [89], [^\_A-Za-z0-9]}

Имеют место следующие тождества

[\_A-Za-z] = [lL] | [uU] | [\_A-KM-TV\_Za-km-tv-z],

[0-9] = [0] | [1-7] | [89],

[^\_A-Za-z0-9] = [^\_A-Za-z0-9],

которые выражают классы разбиения Π1 через классы разбиения П1 ⋂ П2.

Эти тождества выписываются по строкам таблицы.

Классы разбиения Π2 выражаются через классы разбиения П1 ⋂ П2 аналогично, но тождества выписываются по столбцам таблицы:

[0] = [0],

[1-7] = [1-7],

[lL] = [lL],

[uU] = [uU],

[^01-7 lL uU] = [\_A-KM-TV\_Za-km-tv-z] | [89] | [^\_A-Za-z0-9].