Регулярные выражения

Содержание

Назначение регулярных выражений

Регулярные языки, кроме автоматных грамматик и конечных автоматов, часто удобнее описывать в виде некоторого алгебраического языка, который получил название **регулярных выражений**.

На практике регулярные выражения весьма полезная вещь, поэтому библиотеки по работе с регулярными выражениями поддерживаются почти всеми языками программирования. Они позволяют осуществлять очень гибкую обработку текстовых цепочек.

Например, все вы привыкли, что при организации поиска символ $*$ обозначает любую последовательность символов, символ ? - любой один символ. На самом деле, используя эти символы, вы составляете некоторое регулярное выражение, которое по сути является некоторым шаблоном для поиска.

Оказывается, составляя определенный шаблон, или совокупность шаблонов, вы задаете некоторый регулярный язык. Причем, в отличие от порождающих грамматик, конечных автоматов, он задается более декларативно.

Назначение регулярных выражений

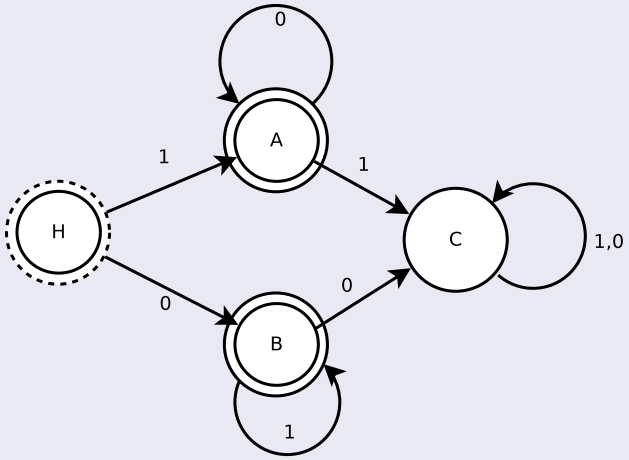
Вот несколько примеров использования регулярных выражений:

- Команды поиска, например, команда grep операционной системы UNIX или аналогичные команды для поиска цепочек, которые можно встретить в текстах.
- Лексические анализаторы (а также специальные программы, выполняющие роль генераторов лексических анализаторов) - это компонент компилятора, разбивающий исходную программу на логические единицы (лексемы), которые состоят из одного или нескольких символов и имеют определенный смысл. Примерами лексем являются ключевые слова (например while), идентификаторы (любая буква, за которой следует нуль или несколько букв и/или цифр) и такие знаки, как + или <=.

В качестве примера рассмотрим также язык $L = \{0(1)^n, 1(0)^n : n = 0, 1, \cdots\}$. Данный язык состоит из цепочек двух типов: вначале один нуль и далее любое количество единиц, или вначале одна единица и далее любое количество нулей.
Для него очень просто построить конечный автомат, распознающий правильные фразы этого языка, по которому можно также легко построить леволинейную грамматику.

Назначение регулярных выражений

Конечный автомат, задающий язык $L = \{0(1)^n, 1(0)^n : n = 0, 1, \cdots\}$



Леволинейная автоматная грамматика, синтезированная по конечному автомату:

$$A \to 1, B \to 0, C \to A1, C \to B0, A \to A0, B \to B1, S \to A|B$$

Легко увидеть, что нетерминал C недостижим из S, поэтому окончательно грамматику языка L можно представить в виде:

$$S \to A|B, A \to A0|1, B \to B1|0$$

Этот язык также можно задать с помощью введенной ранее операции звездочка Клини: $L = 10^*$ или 01^*. Данная форма очень похожа на исходное определение: $L = \{0(1)^n, 1(0)^n : n = 0, 1, \cdots\}$, и по сути является декларативной. Такой способ и есть регулярное выражение. Рассмотрим этот способ в более строгих терминах.

Определение регулярных множеств и выражений

Каждое регулярное выражение по сути является шаблоном, описывающим целое семейство (нередко бесконечное) строк. Прежде чем мы введем строгое определение регулярных выражений, вспомним назначие трех операций, определенных над некоторым алфавитом символов.

1. **Операция объединения:** если A, B - множества некоторых строк, то $A + B$ обозначает множество объединения строк множества A и строк множества B. Если под A понимается некоторый шаблон, которому удовлетворяют, например, строки из множества $\{a, aaa, aba\}$, и B некоторый шаблон, которому удовлетворяют строки из множества $\{bbb, baab, bbba\}$, то $A + B = \{a, aaa, aba, bbb, baab, bbba\}$. Заметим, что если $A = \varnothing$, то $\varnothing + B = B + \varnothing = B$
2. **Операция конкатенации:** если A, B - множества некоторых строк, то $A \cdot B$ обозначает множество строк, которое можно образовать путем дописывания к любой цепочке из A любой цепочки из B. Например, если шаблон $A = \{\epsilon, sd, xcv\}$ и шаблон $B = \{re, w\}$, то $A \cdot B = \{re, w, sdre, sdw, xcvre, xcvw\}$. Обратите внимание, что $A \cdot B \neq B \cdot A$ в общем случае, также $\{\epsilon\} \cdot B = B$
3. **Операция итерации:** если A - некоторое множества строк, то A^* представляет собой множество всех тех строк, которые можно образовать путем конкатенации любого количества строк из A. При этом допускаются повторения, т.е. одна и та же строка из A может быть выбрана для конкатенации более одного раза.

Определение регулярных множеств и выражений

Поскольку идея итерации является более сложной рассмотрим ряд примеров. Более просто говорить об итерации в терминах конкатенации. Обозначим $A^0 = \{\epsilon\}$, $A^n = \overbrace{A \cdot A \cdot \ldots \cdot A}^{n\ раз}$, тогда $A^* = \{\epsilon\} \cup A^1 \cup A^2 \cup A^3 \ldots \cup A^n, n \to \infty$
Например, пусть $A = \{0, 1\}$. Имеем $A^0 = \{\epsilon\}, A^1 = A = \{0, 1\}, A^2 = \{00, 01, 10, 11\}, A^3 = \{0, 1\} \cdot \{00, 01, 10, 11\} = \{000, 001, 010, 011, 100, 101, 110, 111\}$.
Таким образом, получаем:

$$A^* = \{\epsilon, 0, 1, 00, 01, 10, 11, 000, 001, 010, 011, 100, 101, 110, 111, \ldots\}$$

Легко видеть, что количество элементов в каждом A^i равно 2^i, а само множество A^* представляет собой все множество последовательностей из нулей и единиц всех возможных длин.
Рассмотрим еще пример A - это бесконечное множестве всех возможных непустых строк из нулей, тогда $A^0 = \{\epsilon\}, A^1 = A, A^2 = A, \forall n : A^n = A$, тогда $A^* = \{\epsilon\} \cup A$.
В качестве последнего примера рассмотрим $A = \varnothing$, тогда $A^0 = \{\epsilon\}, A^1 = A^2 = \varnothing, \forall n : A^n = \varnothing$. Значит $A^* = \varnothing^* = \{\epsilon\}$.

Определение регулярных множеств

Используя введенные операции, а также обозначения $\{\epsilon\}, \varnothing$, можно ввести важное понятие регулярного множества.

Определение

Регулярным множеством над алфавитом символов Σ являются:

1. *$\varnothing$ - регулярное множество;*
2. *$\{\epsilon\}$ - регулярное множество;*
3. *Если $s \in \Sigma$, то $\{s\}$ - регулярное множество;*
4. *Если A, B - регулярные множества, то $A+B$, $A \cdot B$, $B \cdot A$, A^*, B^* - регулярные множества;*
5. *ничто другое не является регулярным множеством.*

Определение регулярных выражений

Регулярные выражения - это более простой способ обозначения регулярных множеств.

Определение

Если задан алфавит символов Σ, над которым определяются регулярные множества, то:

1. *обозначим символом 0 регулярное выражение, которое задает регулярное множество $\varnothing$;*
2. *обозначим символом ϵ регулярное выражение, которое задает регулярное множество $\{\epsilon\}$;*
3. *обозначим символом s регулярное выражение, которое задает регулярное множество $\{s\}, s \in \Sigma$;*
4. *обозначим $a+b, ab, ba, a^*, b^*$ регулярные выражения, которые задают регулярные множества $A+B, A \cdot B, B \cdot A, A^*, B^*$ соответственно, при условии, что A, B - регулярные множества.*

Свойства регулярных выражений

Два регулярных выражения a, b равны $a = b$, если они задают одно и тоже регулярное множество. При этом каждое регулярное выражение задает только одно регулярное множество, однако одно и тоже регулярное множество можно задавать с помощью разных регулярных выражений.

При записи регулярных выражений можно использовать круглые скобки, указывая приоритет выполнения операций, как в обычных арифметических выражениях. При отсутствии скобок операции выполняются слева направо с учетом их естественных приоритетов:

1. операция итерации имеет наивысший приоритет;
2. операция конкатенации имеет средний приоритет;
3. операция объединения имеет низший приоритет.

Свойства регулярных выражений

Если a, b, c - регулярные выражения, то справедливы следующие формулы:

1. $aa^* = a^*a = a^+$, где a^+ - это a^* без символа ϵ
2. $a + b = b + a, a + (b + c) = (a + b) + c$
3. $a(b + c) = ab + ac, (b + c)a = ba + ca$
4. $a(bc) = (ab)c$
5. $a + a = a, a + a^* = a^*$
6. $\epsilon + a^* = a^*, 0^* = \epsilon$
7. $0\epsilon = \epsilon 0 = 0$
8. $0 + a = a + 0 = a$
9. $\epsilon \cdot a = a\epsilon = a$
10. $(a^*)^* = a^*$
11. $0a = a0 = 0$

Данные свойства можно обосновать, основываясь на свойствах операций над соответсвующими регулярными множествами.

Уравнения с регулярными коэффициентами

Используя данные свойства регулярных выражений, можно проводить их своеобразные преобразования - упрощения:

$$(a^* + b) \cdot a \cdot 0^* = (a^* + b) \cdot a \cdot \epsilon = (a^* + b) \cdot a = a^*a + ba = a^+ + ba$$

С другой стороны,

$$(a^* + b) \cdot a \cdot 0^* = (a^*a + ba) \cdot 0^* = (a^+ + ba)\epsilon = a^+\epsilon + ba\epsilon = a^+ + ba$$

а также:

$$(a^* + b) \cdot a \cdot 0^* = (a^* + b) \cdot 0^* \cdot a = (a^*\epsilon + b\epsilon) \cdot a = (a^* + b) \cdot a = a^*a + ba = a^+ + ba$$

Можно даже составлять уравнения и пытаться их решать. Например, для заданных регулярных выражений a, b требуется найти такие регулярные выражения x_1, x_2, чтобы выполнялись равенства:

$$x_1 = ax_1 + b$$

$$x_2 = x_2a + b$$

Уравнения с регулярными коэффициентами

Для решения уравнения

$$x_1 = ax_1 + b$$

положим $x_1 = cb$, т.к. ясно, что x_1 содержит регулярные выражения b:

$$x_1 = ax_1 + b = (ac + \epsilon)b \Rightarrow c = ac + \epsilon$$

Поскольку в последнем c содержит само себя и a, то $c = a^*$, действительно, $c = ac + \epsilon = aa^* + \epsilon = a^+ + \epsilon = a^*$.
Окончательно, $x_1 = a^*b$
Действительно,

$$x_1 = ax_1 + b = aa^*b + b = (aa^* + \epsilon)b = a^*b$$

Рассуждая аналогично, можно получить, что $x_2 = ba^*$ (убедитесь в этом самостоятельно).

Системы уравнений с регулярными коэффициентами

Кроме отдельных уравнений, можно составлять системы уравнений с регулярными коэффициентами, которые имеют важное значение при синтезе регулярных выражений на основе порождающих регулярных грамматик. Соответственно возможно два вида таких систем уравнений:

Правосторонняя запись:

$$\begin{cases} X_1 = a_{10} + a_{11}X_1 + a_{12}X_2 + \ldots + a_{1n}X_n \\ X_2 = a_{20} + a_{21}X_1 + a_{22}X_2 + \ldots + a_{2n}X_n \\ \quad \ldots \\ X_i = a_{i0} + a_{i1}X_1 + a_{i2}X_2 + \ldots + a_{in}X_n \\ \quad \ldots \\ X_n = a_{n0} + a_{n1}X_1 + a_{n2}X_2 + \ldots + a_{nn}X_n \end{cases}$$

Левосторонняя запись:

$$\begin{cases} X_1 = a_{10} + X_1a_{11} + X_2a_{12} + \ldots + X_na_{1n} \\ X_2 = a_{20} + X_1a_{21} + X_2a_{22} + \ldots + X_na_{2n} \\ \quad \ldots \\ X_i = a_{i0} + X_1a_{i1} + X_2a_{i2} + \ldots + X_na_{in} \\ \quad \ldots \\ X_n = a_{n0} + X_1a_{n1} + X_2a_{n2} + \ldots + X_na_{nn} \end{cases}$$

Алгоритм решения системы уравнений с регулярными коэффициентами на примере правосторонней записи

Данные системы решаются методом последовательных подстановок. Продемонстрируем этот метод на примере системы с правосторонней записью из двух уравнений:

$$\begin{cases} X_1 = a_{10} + a_{11}X_1 + a_{12}X_2 \\ X_2 = a_{20} + a_{21}X_1 + a_{22}X_2 \end{cases}$$

Положим из первого уравнения $X_1 = a_{11}X_1 + b_1, b_1 = a_{10} + a_{12}X_2$, тогда получаем $X_1 = a_{11}^* b_1 = a_{11}^*(a_{10} + a_{12}X_2)$. Подставляем полученное соотношение во второе уравнение:

$$X_2 = a_{20} + a_{21}a_{11}^*(a_{10} + a_{12}X_2) + a_{22}X_2$$

Раскрываем скобки, проводим группировки:

$$X_2 = a_{20} + a_{21}a_{11}^*a_{10} + a_{21}a_{11}^*a_{12}X_2 + a_{22}X_2$$

$$X_2 = (a_{21}a_{11}^*a_{12} + a_{22})\,X_2 + (a_{20} + a_{21}a_{11}^*a_{10})$$

$$X_2 = (a_{21}a_{11}^*a_{12} + a_{22})^*\,(a_{20} + a_{21}a_{11}^*a_{10})$$

$$X_1 = a_{11}^*(a_{10} + a_{12}\,(a_{21}a_{11}^*a_{12} + a_{22})^*\,(a_{20} + a_{21}a_{11}^*a_{10}))$$

Система уравнений с регулярными коэффициентами всегда имеет решение, но это решение не единственное. Предложенный алгоритм всегда находит хотя бы одно такое решение.

Построение системы уравнений с регулярными коэффициентами на основе леволинейной регулярной грамматики

Пусть задана леволинейная регулярная грамматика $N = \{X_1, X_2, \ldots, X_n\}$. Все правила этой грамматики имеют вид: $X_i \to X_j\gamma$, или $X_i \to \gamma$. Целевой символ S будет соответствовать некоторому X_k, который собственно и следует найти, чтобы задать соответствующий язык в виде регулярного выражения.

Строим систему уравнений с регулярными коэффициентами на основе переменных $X_1, X_2, \ldots, X_n$:

$$\begin{cases} X_1 = a_{10} + X_1a_{11} + X_2a_{12} + \ldots + X_na_{1n} \\ X_2 = a_{20} + X_1a_{21} + X_2a_{22} + \ldots + X_na_{2n} \\ \quad \ldots \\ X_i = a_{i0} + X_1a_{i1} + X_2a_{i2} + \ldots + X_na_{in} \\ \quad \ldots \\ X_n = a_{n0} + X_1a_{n1} + X_2a_{n2} + \ldots + X_na_{nn} \end{cases}$$

Коэффициенты в системе уравнений принимаем согласно ряда правил:

Построение системы уравнений с регулярными коэффициентами на основе леволинейной регулярной грамматики

Правила формирования коэффициентов системы уравнений с регулярными коэффициентами:

- Если в множестве правил грамматики есть правила вида $X_i \to \gamma_1|\gamma_2|\ldots|\gamma_m$, то принимаем $a_{i0} = \gamma_1 + \gamma_2 + \ldots + \gamma_m$, иначе $a_{i0} = 0$
- Если в множестве правил грамматики есть правила вида $X_i \to X_j\gamma_1|X_j\gamma_2|\ldots|X_j\gamma_m$, то принимаем $a_{ij} = \gamma_1 + \gamma_2 + \ldots + \gamma_m$, иначе $a_{ij} = 0$

Решив данную систему и найдя X_k соответствующий целевому символу S, мы найдем регулярное выражение, задающее тот же язык, что и регулярная грамматика.

Пример построения регулярного выражения на основе леволинейной регулярной грамматики

Рассмотрим задание языка $L = \{0(1)^n, 1(0)^n : n = 0, 1, \cdots\}$ с помощью регулярного выражения. Как мы выяснили раньше, данный язык можно задать следующей грамматикой $S \to A|B, A \to A0|1, B \to B1|0$

Введем обозначения и воспользуемся рассмотренным алгоритмом: $A = X_1, B = X_2, S = X_3$. Таким образом, найдя решение для X_3, мы получим итоговое регулярное выражение. С учетом введенных обозначений грамматику можно переписать в виде:

$$X_3 \to X_1\epsilon|X_2\epsilon, X_1 \to X_10|1, X_2 \to X_21|0$$

Формируем коэффициенты при X_1. Для получения a_{10} смотрим все правила вида $X_1 \to \gamma_1|\gamma_2|\ldots|\gamma_m$, тогда $a_{10} = \gamma_1 + \ldots + \gamma_m$, у нас всего одно такое правило $X_1 \to 1$, значит $a_{10} =$ "1". Для получения коэффициентв $a_{1j}, j > 0$ ищем правила вида $X_1 \to X_j\gamma_1|X_j\gamma_2|\ldots|X_j\gamma_m$, тогда $a_{ij} = \gamma_1 + \ldots + \gamma_m$. В нашем случае для a_{11} такое правило только одно $X_1 \to X_10$, значит $a_{11} =$ "0". Для a_{12}, a_{13} таких правил нет, значит $a_{12} = 0, a_{13} = 0$ (обратите внимание, в данном случае нуль берется без кавычек, что означает не символ «0», а регулярное выражение, обозначающее регулярное множество $\varnothing$). Аналогичным образом легко получить коэффициенты при X_2, X_3:

для X_2: $a_{20} =$ "0", $a_{21} = 0, a_{22} =$ "1", $a_{23} = 0$

для X_3: $a_{30} = 0, a_{31} = \epsilon, a_{32} = \epsilon, a_{33} = 0$

Пример построения регулярного выражения на основе леволинейной регулярной грамматики

Окончательно получаем следующую систему уравнений с регулярными коэффициентами:

$$\begin{cases} X_1 = \text{"1"} + X_1\text{"0"} + X_20 + X_30 \\ X_2 = \text{"0"} + X_10 + X_2\text{"1"} + X_30 \\ X_3 = 0 + X_1\epsilon + X_2\epsilon + X_30 \end{cases}$$

Используя свойства регулярных выражений: $0a = a0 = 0, 0 + a = a + 0 = a, \epsilon a = a\epsilon = a$, получаем:

$$\begin{cases} X_1 = \text{"1"} + X_1\text{"0"} \\ X_2 = \text{"0"} + X_2\text{"1"} \\ X_3 = X_1 + X_2 \end{cases}$$

Используя рассмотренное решение $x = xa + b \Rightarrow x = ba^*$, получаем:

$$X_1 = \text{"1""0"}^*$$

$$X_2 = \text{"0""1"}^*$$

$$X_3 = \text{"1""0"}^* + \text{"0""1"}^*$$

Полученное регулярное выражение для X_3, как легко убедиться, действительно задает язык $L = \{0(1)^n, 1(0)^n : n = 0, 1, \cdots\}$.

Задание 1

Решите систему уравнений с регулярными коэффициентами:

$$\begin{cases} X_1 = \text{"0"}X_2 + \text{"1"}X_1 \\ X_2 = \text{"0"}X_3 + \text{"1"}X_2 \\ X_3 = \text{"0"}X_1 + \text{"1"}X_3 \end{cases}$$

Задание 2

Постройте регулярную грамматику, которая описывает язык идентификаторов из одной буквы «a» и цифр 0, 1. Запрещается начинать идентификатор с цифры. Например, разрешенными являются следующие идентификаторы: «aaa», «a0», «aa10», «a1a0aa» и т.д. Постройте и решите систему уравнений с регулярными коэффициентами на основе этой грамматики.