**Регулярные выражения и регулярные множества**

Регулярные выражения — альтернативный, отличный от порождающих грамматик и синтаксических диаграмм и имеющий свои преимущества, способ задания языка.

Регулярное выражение обозначает (порождает) множество цепочек, которое называют регулярным множеством.

Множество цепочек, соответствующее регулярному выражению R, будем обозначать R^.

Регулярные выражения над алфавитом Σ образуются по следующим правилам:

1. Отдельный символ алфавита а∈Σ является регулярным выражением. Обозначаемое таким выражением множество цепочек есть {а}, то есть состоит из одной цепочки а.

2. Пустая цепочка ε есть регулярное выражение. Обозначает регулярное множество {ε}.

3. Если R и Q — регулярные выражения над алфавитом Σ, то запись RQ (конкатенация) также является регулярным выражением. Множество, обозначаемое RQ, состоит из всех цепочек, образованных конкатенацией двух цепочек, так что первая цепочка пары порождается выражением R а вторая — выражением Q

Формально это может быть записано таким образом: (RQ)^ = {αβ} | α ∈R^, β ∈Q^}.

4. Если R и Q — регулярные выражения над алфавитом Σ, то запись R | Q (читается «R или Q») также является регулярным выражением и обозначает регулярное множество R^ ∪ Q^, то есть множество всех цепочек, порождаемых как выражением R, так и выражением Q.

5. Если R — регулярное выражение над алфавитом Σ, то запись R\* (итерация R) также является регулярным выражением и обозначает множество всех цепочек, полученных повторением цепочек, порождаемых R, ноль или более раз.

6. Если R — регулярное выражение над алфавитом Σ, то (R) (R в скобках) также является регулярным выражением, которое обозначает то же множество, что и R.

Предполагается определенный приоритет операций, с помощью которых образуются регулярные выражения.

Наивысший приоритет имеет итерация (знак «\*»), далее — конкатенация, далее — «или» (знак «|»). Скобки используются для изменения порядка операций.

*Пример 1*. С помощью регулярного выражения можно задать правила записи целых чисел со знаком:

(+ | -| ε) *цц*\*

где *ц* обозначает любую цифру от 0 до 9.

Если это не вызывает разночтения, символ ε можно не записывать.

Повторение один или более раз иногда обозначают знаком «+». R+=RR\* Другая форма выражения, определяющего целые:

(+|-)*ц*+.

Нетрудно, впрочем, записать выражение, обозначающее множество всех целых со знаком, не прибегая к условному обозначению цифр с помощью «*ц*»:

(+|-)(0|1|2|3|4|5|6|7|8|9)+

*Пример 2*. Регулярное выражение, задающее множество идентификаторов:

*б*(*б*|*ц*)\*

где *б* — буква; *ц* — цифра.

**Эквивалентность регулярных выражений и автоматных грамматик**

Автоматные языки являются регулярными множествами. Регулярные множества являются автоматными языками.

Сказанное означает, что для любой автоматной грамматики можно записать такое регулярное выражение, что обозначаемое этим выражением множество цепочек совпадает с языком, порождаемым грамматикой.

И, наоборот, для любого регулярного выражения можно найти автоматную грамматику, порождающую то же множество цепочек, что и регулярное выражение.

Будем считать, что автоматный язык задается синтаксической диаграммой. Можно установить взаимно однозначное соответствие между конструкциями, из которых строятся регулярные выражения (правила 1-6) и фрагментами, из которых состоят синтаксические диаграммы автоматных грамматик.

Это соответствие показано в табл. 2.

*Таблица 2. Эквивалентность регулярных выражений и автоматных грамматик*

|  |  |  |
| --- | --- | --- |
| Номер правила | Фрагмент выражения | Участок диаграммы |
| 1 | a |  |
| 2 | ε | \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |
| 3 | RQ |  |
| 4 | R|Q |  |
| 5 | R\* |  |
| 6 | (R) |  |

Используя такое соответствие, по выражению можно построить диаграмму, а по диаграмме — регулярное выражение. Уточнение деталей таких построений, которым мы не будем здесь заниматься, и доказывает справедливость сформулированного ранее утверждения об эквивалентности.

Уместно напомнить, что *автоматные грамматики* называют также *регулярными*.

**Для чего нужны регулярные выражения**

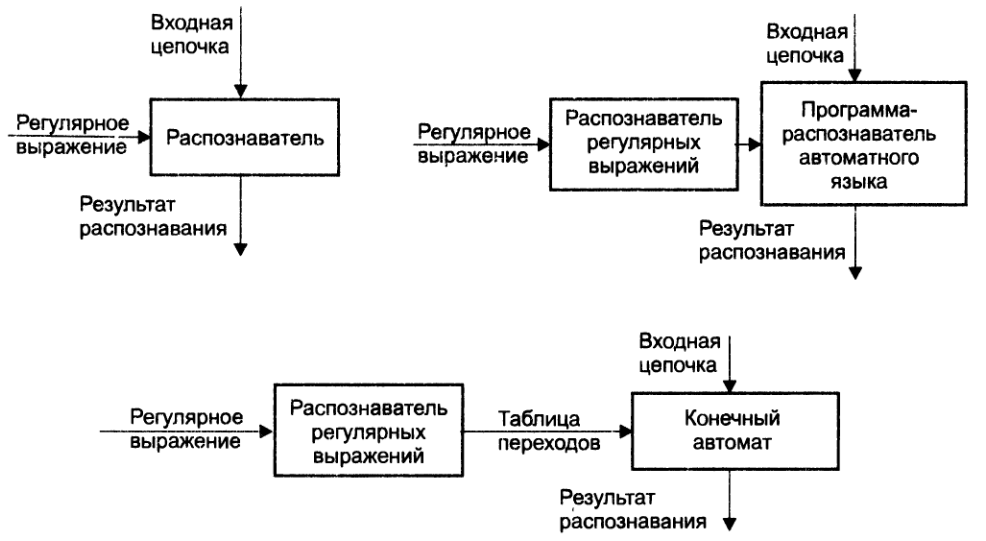
Автоматные грамматики, регулярные выражения и синтаксические диаграммы являются эквивалентными способами задания автоматных языков.

По сравнению с грамматиками синтаксические диаграммы обладают большей наглядностью.

Регулярные же выражения имеют то важное достоинство, что представляют собой строки символов, которые могут быть легко обработаны с помощью компьютерных программ.

Обработка регулярного выражения, выступающего в роли исходных данных для некоторой программы, может иметь целью его анализ, преобразование и даже... создание распознавателя автоматного языка, порождаемого этим выражением.

Последнее представляет безусловный интерес, поскольку открывает возможность автоматизации построения синтаксических анализаторов. Работа подобной программы может происходить по одной из схем, показанных на рис. 16.



*Рис.16. Использование регулярных выражений*

Регулярные выражения наглядней порождающих грамматик. Это обусловлено тем, что предусмотрено явное обозначение повторения (знак итерации «\*»). В нотации грамматик итерация задается с помощью рекурсии.

Сравните, например, грамматики G10 и G12, задающие язык идентификаторов, с эквивалентным регулярным выражением из *примера 2*.

**Регулярные выражения как языки**

Регулярное выражение над алфавитом Σ — это цепочка символов в расширенном алфавите

Σ ∪ {|, \*, (, ) }.

Множество всех таких цепочек-выражений образует язык.

Возникает естественный вопрос, к языкам какого типа по классификации Н. Хомского этот язык принадлежит.

К автоматным? Тогда, быть может, правила записи регулярных выражений можно задать регулярным выражением? Нет, нельзя.

Синтаксис регулярных выражений может быть определен только контекстно-свободной, но не автоматной грамматикой. Вот эта грамматика:

R→a | RR | R\* | R" | "R | (R) | ε

В этой записи есть ряд условностей:

а обозначает любой символ алфавита Σ,

запись "|", представляет знак «|», используемый в регулярных выражениях и совпадающий с аналогичным знаком, применяемым при записи грамматик.

Приведенная грамматика не отражает принятый для регулярных выражений порядок операций. Грамматика, трактующая структуру регулярного выражения в соответствии с приоритетами операций, может быть записана так:

R→ T | R" | "T

T→ M | RM

М→ a | М\* | (R) | ε.

**Расширенная нотация для регулярных выражений**

Регулярные выражения - это строки символов, и тем они интересны как средство задания автоматных языков.

Но использование надстрочных знаков «\*» и «+» несколько затрудняет запись выражений и их считывание компьютерной программой. Получили распространение другие варианты обозначений.

Повторение ноль или более раз обозначают фигурными скобками:

R\* = {R}.

Используются также квадратные скобки, обозначающие необязательность заключенного в них выражения:

[R] = (R|ε).

Знаки «\*» и «+» в этом случае уже не используются.

Соглашения о способах записи символов, с помощью которых строятся сами выражения (скобки, знак «|»), в случае, если они также входят в терминальный алфавит, могут быть разными.

Можно заключать такие метасимволы в кавычки «"». При необходимости записать саму кавычку ее заключают в апострофы «'», а апостроф, если нужно, записывается в кавычках.

По этим правилам регулярные выражения, обозначающие множество целых со знаком и множество идентификаторов, будут выглядеть так:

[ +|- ] *ц* { *ц* }

*б* { *б*|*ц* }.

На этом мы заканчиваем рассмотрение автоматных грамматик, в ходе которого удалось построить простые и эффективные методы распознавания автоматных языков.

С помощью автоматных грамматик определяется синтаксис простейших элементов языков программирования: идентификаторов, чисел, других констант, знаков операций и разделителей.