Регулярные выражения

Истоки регулярных выражений лежат в теории автоматов, теории формальных языков и классификации формальных грамматик.

В 40-х годах прошлого столетия нервная система была описана используя простой автомат в качестве модели нейрона.

Позже ученный-математик описал эти модели используя свою систему математических обозначений, названную «регулярные множества».

Впервые регулярные выражения стали использоватся в UNIX (GREP).

Сегодня регулярные выражения используются во многих современных инструментах, таких как PHP и Apache. Так же используются в веб-программировании, системном администрировании и т. п.

Регулярные выражения

Регулярные выражения — аналитический или формульный способ задания регулярных языков. Они состоят из констант и операторов, которые определяют множества строк и множество операций над ними.

Пусть будет Σ — конечный алфавит. Определим следующие операции

+ - объединение,

$\bullet$ - конкатенация,

* - итерация

(, *)* скобки

при помощи которых можно составлять выражения с элементами из Σ.

Если $a, b \in \Sigma$ тогда

$a \bullet b \in \Sigma$

$a+b \in \Sigma$

$a^*=\{\varepsilon, a, aa, aaa, \dots\} \in \Sigma$,

$b^*=\{\varepsilon, b, bb, bbb, \dots\} \in \Sigma$.

Регулярные выражения

Определение: **Регулярным выражением** над алфавитом **Σ**, называется выражение построенное рекурсивно следующим образом:

1) $\varnothing$ (пустой/нулевой язык) является регулярным выражением;

2) **ε** является регулярным выражением;

3) Если **a**$\in$**Σ**, то **a** является регулярным выражением;

4) Если **r** и **s** являются регулярными выражениями, то **r+s**, **r·s**, **(r)**, **(s)**, **r***, **s*** тоже являются регулярными выражениями;

5) Других регулярных выражений нет.

(Для экономии вместо **r·s** часто пишут просто **rs**).

Регулярные выражения

Обозначим множество регулярных выражений над алфавитом Σ как $R(\Sigma)$. И пусть будет $e \in R(\Sigma)$.

Тогда каждое регулярное выражение ***e*** *задаёт* некоторый язык над алфавитом Σ (обозначение **L(e)**), определяемый рекурсивно следующим образом:

- Если $e=\varnothing \Rightarrow$ L(e)= $\varnothing$,
- Если $e=\varepsilon \Rightarrow$ L(e)={ε},
- Если $e=a \in \Sigma \Rightarrow$ L(e)= L(a)={a},
- Если $e=a+b \Rightarrow$ L(e)= $L(a) \cup$ L(b),
- Если $e=a \bullet b \in \Sigma \Rightarrow$ L(e)= $L(a) \bullet$ L(b),
- Если $e=a^* \Rightarrow$ L(e)= (L(a))*.

Примечание: Часто вместо ***L(e)*** часто пишут просто ***e***.

Регулярные выражения

Примеры:

Регулярное выражение	Значение регулярного выражения
01	единственная строка 01
0+1	две строки: 0 и 1
1^*	строки, образованные из единиц, включая пустую строку
$(0+1)^*$	строки, образованные из символов 0 и 1, включая пустую строку
$0+1^*$	строки, состоящие из нуля и любой строки единиц, включая пустую
$(0+1)^*011$	строки, образованные из символов 0 и 1, включая пустую, обязательно оканчивающиеся строкой 011

Регулярные выражения

Примеры:

1) Регулярное выражение $(0+1)^*$ представляет множество всех слов в алфавите {0,1}: ε, 0, 1, 01, 00001, 1010101010 и т. д.

$L((0+1)^*)=(L(0+1))^*=(L(0)+L(1))^*=(\{0\}\cup\{1\})^*=\{0+1\}^*$

2) Регулярное выражение $11(0+1)^*001$ представляет язык состоящий из всех слов в алфавите {0,1}, начинающихся с 11, заканчивающихся на 001.

Регулярные выражения

<u>Определение</u>: Язык L называется **регулярным**, если он задаётся некоторым регулярным выражением.

Теорема: Язык L, определённый над алфавитом Σ, является регулярным тогда и только тогда, когда существует $e \in R(\Sigma)$ и $L(e)=L$.

Регулярные выражения

Свойства регулярных выражений

Лемма: Для любых регулярных выражений e, f, g выполняются следующие тождества:

1. $e+f=f+e$
2. $e+\varepsilon=\varepsilon+e=e$
3. $(e+f)+g=e+(f+g)$ - транзитивность
4. $e\cdot\varepsilon=\varepsilon\cdot e=e$
5. $(e\cdot f)\cdot g=e\cdot(f\cdot g)$
6. $e\cdot(f+g)=e\cdot f+e\cdot g$ - дистрибутивность
7. $(f+g)\cdot e=f\cdot e+g\cdot e$
8. $f\cdot g \neq g\cdot f$
9. $(e^*)^*=e^*$

Последовательность выполнения операций: (), *, ·, +

Примечание: Равенство понимается как равенство языков, задаваемых регулярными выражениями.

Регулярные выражения

Теорема: По каждому детерминированному или недетерминированному конечному автомату можно построить регулярное выражение, которое представляет язык, распознаваемый этим автоматом.

Теорема: Классы языков, задаваемые детерминированными и недетерминированными конечными автоматами, регулярными выражениями совпадают.

Т. е. класс автоматных языков и класс регулярных языков совпадают.

Регулярные выражения

Выводы:

Регулярное выражение это формальный язык поиска и осуществления манипуляций с подстроками в тексте, основанный на использование метасимволов.

Многие современные языки программирования имеют встроенную поддержку регулярых выражений: PERL, Java, PHP, JavaScript. Некоторые текстовые редакторы и СУБД используют регулярные выражения для поиска и подстановки текста. С помощью регулярных выражений можно:

- Проверить соответствует ли вся строка целиком заданному шаблону;
- Найти в строке подстроки, удовлетворяющие заданному шаблону;
- Извлечь из строки подстроки, соответсвующие заданному шаблону;
- Изменить в строке подстроки соответсвующие шаблону.

Регулярные выражения

Задания:

Написать регулярное выражения которое задает множества всех слов из букв ***a, b, c***, в которых слово ***bac*** является подсловом. Построить КА.

Построить КА для следующих регулярных выражений:

- 0*11*
- 01*0(0+1)*
- (a+b)* и слова заканчиваются на aaa или abb
- aa(a+b)*b
- b*aa*b(a+b)*

Регулярные выражения