Регулярные выражения

Регуля́рные выраже́ния (англ. *regular expressions*, жарг. ***регэ́кспы*** или ***ре́гексы***) — система обработки текста, основанная на специальной системе записи образцов для поиска. Образец (англ. *pattern*), задающий правило поиска, по-русски также иногда называют «шаблоном», «маской».

Сейчас регулярные выражения используются многими текстовыми редакторами и утилитами для поиска и изменения текста на основе выбранных правил. Многие языки программирования уже поддерживают регулярные выражения для работы со строками. Например, Perl и Tcl имеют встроенный в их синтаксис механизм обработки регулярных выражений. Набор утилит (включая редактор sed и фильтр grep), поставляемых в дистрибутивах Unix, одним из первых способствовал популяризации понятия регулярных выражений.

1. Базовые понятия

Регулярные выражения используются для сжатого описания некоторого множества строк с помощью шаблонов, без необходимости перечисления всех элементов этого множества. При составлении шаблонов используется специальный синтаксис, поддерживающий, обычно, следующие операции:

Перечисление Вертикальная черта разделяет допустимые варианты. Например, «gray|grey» соответствует *gray* или *grey*.

Группировка Круглые скобки используются для определения области действия и приоритета операторов. Например, «gray|grey» и «gr(a|e)y» являются разными образцами, но они оба описывают множество, содержащее *gray* и *grey*.

Квантификация *Квантификатор* после символа или группы определяет, сколько раз предшествующее выражение может встречаться.

{m,n} общее выражение, повторений может быть **от m до n включительно**.

{m,} общее выражение, **m и более повторений**.

{,n} общее выражение, **не более n повторений**.

? Знак вопроса означает **0 или 1** раз, то же самое, что и **{0,1}**. Например, «colou?r» соответствует и *color*, и *colour*.

***** Звёздочка означает **0, 1 или любое число** раз (**{0,}**). Например, «go*gle» соответствует *ggle*, *gogle*, *google* и др.

+ Плюс означает **хотя бы 1** раз (**{1,}**). Например, «go+gle» соответствует *gogle*, *google* и т. д. (но не *ggle*).

Конкретный синтаксис регулярных выражений зависит от реализации.

2. В теории формальных языков

Регулярные выражения состоят из констант и операторов, которые определяют множества строк и множества операций на них соответственно. На данном конечном алфавите Σ определены следующие константы:

- (*пустое множество*) $\varnothing$ обозначает $\varnothing$
- (*пустая строка*) ε обозначает множество $\{\varepsilon\}$
- (*строка*) a в Σ обозначает множество $\{a\}$

и следующие операции:

- (*связь*, *конкатенация*) RS обозначает множество $\{\ \alpha\beta \mid \alpha$ из R и β из $S\ \}$. Пример: {"ab", "c"}{"d", "ef"} = {"abd", "abef", "cd", "cef"}.
- (*перечисление*) $R|S$ обозначает объединение R и S.
- (*замыкание Клини, звезда Клини*) R^* обозначает минимальное супермножество из R, которое содержит ε и закрыто связью строк. Это есть множество всех строк, которые могут быть получены связью нуля или более строк из R. Например, {"ab", "c"}* = {ε, "ab", "c", "abab", "abc", "cab", "cc", "ababab", … }.

Многие книги используют символы $\cup$, + или $\vee$ для перечисления вместо вертикальной черты.

3. Синтаксис

3.1. Традиционные регулярные выражения в Unix

Синтаксис «базовых» регулярных выражений Unix на данный момент определён POSIX как устаревший, но он до сих пор широко распространён из соображений обратной совместимости. Многие Unix-утилиты используют такие регулярные выражения по умолчанию.

В этом синтаксисе большинство символов соответствуют сами себе («a» соответствует «a» и т. д.). Исключения из этого правила называются метасимволами:

Различные реализации регулярных выражений интерпретируют обратную косую черту перед метасимволами по-разному. Например, egrep и Perl интерпретируют скобки и вертикальную черту как метасимволы, если перед ними *нет* обратной косой черты и воспринимают их как обычные символы, если черта есть.

Многие диапазоны символов зависят от выбранных настроек локализации. POSIX стандартизовал объявление некоторых классов и категорий символов, как показано в следующей таблице:

3.2. Защита метасимволов

Способ представить сами метасимволы ., - [ ] и другие в регулярных выражениях без интерпретации, то есть, в качестве простых (не специальных) символов — предварить их обратной косой чертой: \. Например, чтобы представить сам символ «точка» (просто точка, и ничего более), надо написать \. (обратная косая черта, а за ней - точка). Чтобы представить символ открывающей квадратной скобки [, надо написать \[(обратная косая черта и следом за ней скобка [) и т.д. Сам метасимвол \ тоже может быть защищен, то есть представлен как \\ (две обратных косых черты), и тогда интерпретатор регулярных выражений воспримет его как простой символ обратной косой черты \.

3.3. «Жадные» выражения

Квантификаторам в регулярных выражениях соответствует максимально длинная строка из возможных (квантификаторы являются «жадными», англ. *greedy*). Это может оказаться значительной проблемой. Например, часто ожидают, что выражение (<.*>) найдёт в тексте теги HTML. Однако этому выражению соответствует целиком строка

Википедия — свободная энциклопедия, в которой каждый может изменить или дополнить любую статью .

Эту проблему можно решить двумя способами. Первый состоит в том, что в регулярном выражении учитываются символы, *не* соответствующие желаемому образцу (<[^>]*> для вышеописанного случая). Второй заключается в определении квантификатора как нежадного (ленивого, англ. *lazy*)— большинство реализаций позволяют это сделать, добавив после него знак вопроса.

Например, выражению (<.*?>) соответствует не вся показанная выше строка, а отдельные теги (выделены цветом):

Википедия — свободная энциклопедия, в которой каждый может изменить или дополнить любую статью

- *? - «не жадный» («ленивый») эквивалент *
- +? - «не жадный» («ленивый») эквивалент +
- {n,}? - «не жадный» («ленивый») эквивалент {n,}

Использование «ленивых» квантификаторов, правда, может повлечь за собой обратную проблему, когда выражению соответствует слишком короткая строка.

Так же существуют квантификаторы повышения жадности, то что захвачено ими однажды назад уже не отдается. Сверхжадные квантификаторы (possessive quantifiers)

- *+ - «сверхжадный» эквивалент *
- ++ - «сверхжадный» эквивалент +
- {n,}+ - «сверхжадный» эквивалент {n,}

3.4. Современные (расширенные) регулярные выражения в POSIX

Регулярные выражения в POSIX аналогичны традиционному Unix-синтаксису, но с добавлением некоторых метасимволов:

Также было отменено использование обратной косой черты: \{…\} становится {…} и \(…\) становится (…).

3.5. Perl-совместимые регулярные выражения (PCRE)

Регулярные выражения в Perl имеют более богатый и в то же время предсказуемый синтаксис, чем даже в POSIX. По этой причине очень многие приложения используют именно Perl-совместимый синтаксис регулярных выражений.

3.6. Группы

( ) Простая группа с захватом.

(?:) Группа без захвата. То же самое, но заключённое в скобках выражение не добавляется к списку захваченных фрагментов. Например, если требуется найти или «здравствуйте», или «здрасте», но не важно, какое именно приветствие найдено, можно воспользоваться выражением здра(?:сте|вствуйте).

(?=) Группа с положительной опережающей проверкой (positive lookahead assertion). Продолжает поиск только если справа от текущей позиции в тексте находится заключённое в скобки выражение. При этом само выражение не захватывается. Например, говор(?=ить) найдёт «говор» в «говорить», но не в «говорит».

(?!) Группа с негативной опережающей проверкой (negative lookahead assertion). Продолжает поиск только если справа от текущей позиции в тексте не находится заключённое в скобки выражение. При этом само выражение не захватывается. Например, говор(?!ить) найдёт «говор» в «говорит», но не в «говорить».

(?<=) Группа с положительной ретроспективной проверкой (positive lookbehind assertion). Продолжает поиск только если слева от текущей позиции в тексте находится заключённое в скобки выражение. При этом само выражение не захватывается. Например, (?<=об)говорить найдёт «говорить» в «обговорить», но не в «уговорить».

(?<!) Группа с отрицательной ретроспективной проверкой (negative lookbehind assertion). Продолжает поиск только если слева от текущей позиции в тексте не находится заключённое в скобки выражение. При этом само выражение не захватывается. Например, (?<!об)говорить найдёт «говорить» в «уговорить», но не в «обговорить».

...

4. Реализации

- NFA (Nondeterministic Finite State Machine; Недетерминированные Конечные Автоматы) используют «жадный» алгоритм отката, проверяя все возможные расширения регулярного выражения в определённом порядке и выбирая первое подходящее значение. NFA может обрабатывать подвыражения и обратные ссылки. Но из-за алгоритма отката традиционный NFA может проверять одно и то же место несколько раз, что отрицательно сказывается на скорости работы. Поскольку традиционный NFA принимает первое найденное соответствие, он может и не найти самое длинное из вхождений (этого требует стандарт POSIX, и существуют модификации NFA выполняющие это требование — GNU sed). Именно такой механизм регулярных выражений используется, например, в Perl, Tcl и .NET.

- DFA (Deterministic Finite-state Automaton; Детерминированные Конечные Автоматы) работают линейно по времени, поскольку не используют откаты и никогда не проверяют какую-либо часть текста дважды. Они могут гарантированно найти самую длинную строку из возможных. DFA содержит только конечное состояние, следовательно, не обрабатывает обратных ссылок, а также не поддерживает конструкций с явным расширением, то есть не способен обработать и подвыражения. DFA используется, например, в lex и egrep.

5. Литература

- *Билл Смит* **Методы и алгоритмы вычислений на строках (regexp)** = Computing Patterns in Strings. — М.: «Вильямс», 2006. — С. 496. — ISBN 0-201-39839-7

- Фридл Дж. Регулярные выражения. Библиотека программиста. — СПб.: Питер, 2001. 352 с. ISBN 5-318-00056-8.

- *Бен Форта* **Освой самостоятельно регулярные выражения (regexp). PHP, Perl, JavaScript, Java, C#(си шарп), Visual Basic, ASP.NET,JSP, MySQL, Unix, Linux** = Sams Teach Yourself Regular Expressions in 10 Minutes. — М.: «Вильямс», 2004. — С. 192. — ISBN 0-672-32566-7

- *Царьков В. Б.* **Теория и методика построения регулярных выражений. Проблема самообразования**. — 2010.

- Ян Гойвертс, Стивен Левитан Регулярные выражения. Сборник рецептов — СПб.: Символ-Плюс, 2010.352 с. ISBN 978-5-93286-181-3, ISBN 978-0-596-52068-7 (англ.)

6. Ссылки

- Практическое использование регулярных выражений для веб-программирования

- Документация, примеры и конструктор регулярных выражений

- Программа JavaScript для тестирования регулярных выражений
- Регулярные выражения в JavaScript
- Регулярные выражения в C#
- Проверка регулярных выражений
- Площадка для тестирования и хранения регулярных выражений
- Онлайн проверка и оптимизация регулярных выражений

7. Источники текстов и изображения, авторы и лицензии

7.1. Текст

- **Регулярные выражения** *Источник:* http://ru.wikibooks.org/wiki/%D0%A0%D0%B5%D0%B3%D1%83%D0%BB%D1%8F%D1%80%D0%BD%D1%8B%D0%B5%20%D0%B2%D1%8B%D1%80%D0%B0%D0%B6%D0%B5%D0%BD%D0%B8%D1%8F?oldid=106230 *Авторы:* MaxSem, Ivan Shmakov, Vl123, Ntfs.hard и Аноним: 25

7.2. Изображения

7.3. Лицензия

- Creative Commons Attribution-Share Alike 3.0