Pontifícia Universidade Católica
do Rio de Janeiro

Marcelo Oikawa

Conversão de regexes para Parsing Expression Grammars

Dissertação de Mestrado

Dissertação apresentada como requisito parcial para obtenção do grau de Mestre pelo Programa de Pós–graduação em Informática do Departamento de Informática da PUC–Rio

Orientador: Prof. Roberto Ierusalimschy

Rio de Janeiro
Agosto de 2010

Pontifícia Universidade Católica
do Rio de Janeiro

Marcelo Oikawa

Conversão de regexes para Parsing Expression Grammars

Dissertação apresentada como requisito parcial para obtenção do grau de Mestre pelo Programa de Pós–graduação em Informática do Departamento de Informática do Centro Técnico Científico da PUC–Rio. Aprovada pela Comissão Examinadora abaixo assinada.

Prof. Roberto Ierusalimschy
Orientador
Departamento de Informática — PUC–Rio

Prof. Luiz Henrique de Figueiredo
IMPA

Prof. Fabio Mascarenhas de Queiroz
Departamento de Informática — PUC-Rio

Prof. José Eugenio Leal
Coordenador Setorial do Centro Técnico Científico — PUC–Rio

Rio de Janeiro, 25 de Agosto de 2010

Marcelo Oikawa

Graduou-se em Ciência da Computação pela Universidade Federal de Viçosa (Viçosa - Minas Gerais).

Ficha Catalográfica

Oikawa, Marcelo

Conversão de regexes para Parsing Expression Grammars / Marcelo Oikawa; orientador: Roberto Ierusalimschy. — 2010

v., 71 f: il. ; 29,7 cm

1. Dissertação (Mestrado em Informática) - Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, 2010

Inclui bibliografia

1. Informática – Teses. 2. Expressões regulares. 3; Regexes; 4. Gramáticas de expressões de parsing; 5. Reconhecimento de linguagens. I. Ierusalimschy, Roberto. II. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Informática. III. Título.

CDD: 004

Agradecimentos

Ao meu orientador, Professor Roberto Ierusalimschy, por toda confiança, dedicação e paciência, além do imenso aprendizado que obtive durante o tempo em que trabalhamos juntos.

A minha grande amiga e 'mãe' Ana Lúcia de Moura por todas as conversas que me confortaram nas horas difíceis.

Aos amigos Thiago Valente e Vinícius Lopes por todo o apoio que me deram desde que cheguei ao Rio.

Aos grandes companheiros do LabLua, Sérgio Madeiros, Fábio Mascarenhas e Chico Sant'Anna, pelas várias horas de conversa e descontração.

A minha família pelo amor incondicional que sempre me fez ter forças para superar qualquer obstáculo.

Ao tratante do Lourival que prometeu uma garrafa de Whisky há um ano atrás e até hoje nada.

Meus sinceros agradecimentos a todos que um dia disseram 'Japa, você é bom e vai conseguir'.

Resumo

Oikawa, Marcelo; Ierusalimschy, Roberto. **Conversão de regexes para Parsing Expression Grammars**. Rio de Janeiro, 2010. 71p. Dissertação de Mestrado — Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

Expressões regulares são um formalismo utilizado para descrever linguagens regulares e compõem a base de diversas bibliotecas de casamento de padrão. No entanto, existem determinados padrões úteis que são complexos ou impossíveis de serem descritos com expressões regulares puras. Devido a essas limitações, linguagens de script modernas disponibilizam bibliotecas de casamento de padrões baseadas em *regexes*, isto é, extensões de expressões regulares compostas, principalmente, por construções ad-hoc que focam em problemas específicos. Apesar de serem muito úteis na prática, os regexes possuem implementações complexas e distantes do formalismo original de expressões regulares. Parsing Expression Grammars (PEG) são uma alternativa formal para reconhecer padrões e possuem mais expressividade que expressões regulares sem necessitar de contruções ad-hoc. O objetivo deste trabalho é estudar formas de conversão de regexes para PEGs. Para isso, estudamos as implementações atuais de regexes e mostramos a conversão de algumas construções para PEGs. Por fim, apresentamos uma implementação da conversão de regexes para PEGs para a linguagem Lua.

Palavras–chave

Expressões regulares; Regexes; Gramáticas de expressões de parsing; Reconhecimento de linguagens.

Abstract

Oikawa, Marcelo; Ierusalimschy, Roberto(Advisor). **Converting regexes to PEGs**. Rio de Janeiro, 2010. 71p. M.Sc Dissertation — Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

Regular expressions are a formalism used to describe regular languages and form the basis of several pattern-matching libraries. However, many interesting patterns either are difficult to describe or cannot be described by pure regular expressions. Because of these limitations, modern scripting languages have pattern matching libraries based on regexes, ie, extensions of regular expressions mainly composed by a set of ad-hoc constructions that focus on specific problems. Although very useful in practice, these implementations are complex and distant from the original formalism of regular expressions. Parsing Expression Grammars (PEG) are a formal alternative to recognize patterns and it is much more expressive than pure regular expressions and does not need use ad-hoc constructions. The goal of this work is to study the convertion of regexes to PEGs. To accomplish this task, we studied the current implementations of regexes and show how to convert some constructions to PEGs. Finally, we present an implementation that convert regexes to PEGs for the Lua language.

Keywords

Regular expressions; Regexes; Parsing Expression Grammars; Theory of Parsing.

Sumário

1 Introdução **11**

2 Regexes **14**
2.1 Expressão independente 16
2.2 Quantificadores 16
2.3 Capturas 18
2.4 Backreferences 19
2.5 Âncoras 20
2.6 Lookahead 21
2.7 Lookbehind 22
2.8 Classes de caracteres 23
2.9 Outras construções 25

3 Parsing Expression Grammars **28**
3.1 LPeg - PEGs em Lua 33

4 Convertendo regexes em PEGs **35**
4.1 Continuation-based conversion 36
4.2 Tamanho da PEG resultante 38
4.3 Regexes $\rightarrow$ PEGs 40
4.4 Capturas 45
4.5 Backreferences 49
4.6 Lookbehind 50
4.7 Âncoras de início 52

5 Lua Regex **54**
5.1 Módulo regex 54
5.2 Módulo lregex 57
5.3 Análise de desempenho 58

6 Conclusão **67**

Lista de figuras

4.1 Backtracking global 40
4.2 Backtracking local 40

Lista de tabelas

2.1 Expressões regulares 14
2.2 Construções mais comuns entre os regexes 15
2.3 Quantificadores 17
2.4 Quantificadores de Lua 18
2.5 Âncoras 21
2.6 Classes pré-definidas de Perl 24
2.7 Classes pré-definidas de Lua 24
2.8 Classes de POSIX 25
2.9 Outras construções 26

3.1 Operadores de PEGs 29

4.1 Tabela de índices 48
4.2 Tabela de capturas 49

5.1 Tempo (em milisegundos) para buscar uma palavra na Bíblia 58
5.2 Tempo (em milisegundos) para casar expressões com repetições 60
5.3 Testes de busca e repetições de `lregex` 62
5.4 Expressões com sequências de alternativas (em milisegundos) 63
5.5 Expressões que buscam duas palavras na mesma frase (em milisegundos) 64
5.6 Expressões que buscam frases com duas palavras especificas (em segundos) 65

Quando nasci, um anjo torto
desses que vivem na sombra
disse: Vai, Carlos! ser gauche na vida.

As casas espiam os homens
que correm atrás de mulheres.
A tarde talvez fosse azul,
não houvesse tantos desejos.

O bonde passa cheio de pernas:
pernas brancas pretas amarelas.
Para que tanta perna, meu Deus, pergunta
meu coração.
Porém meus olhos
não perguntam nada.

O homem atrás do bigode
é serio, simples e forte.
Quase não conversa.
Tem poucos, raros amigos
o homem atrás dos óculos e do bigode.

Meu Deus, por que me abandonaste
se sabias que eu não era Deus
se sabias que eu era fraco.

Mundo mundo vasto mundo,
se eu me chamasse Raimundo
seria uma rima, não seria uma solução.
Mundo mundo vasto mundo,
mais vasto é meu coração.

Eu não devia te dizer
mas essa lua
mas esse conhaque
botam a gente comovido como o diabo.

Carlos Drummond de Andrade