Principles of Programming Languages

COMP251: Syntax and Grammars

Prof. Dekai Wu

Department of Computer Science and Engineering
The Hong Kong University of Science and Technology
Hong Kong, China

Fall 2006

Part III

Regular Grammar, Regular Expression

Regular Grammars

Regular Grammars are a subset of CFGs in which all productions are in one of the following forms:

1. Right-Regular Grammar

```
<A> ::= x
<A> ::= x<B>
```

2. Left-Regular Grammar

```
<A> ::= x
<A> ::= <B>x
```

where A and B are non-terminals and x is a string of terminals.

RE Example 1: Right-Regular Grammar

```
<S> ::= a<A>
<S> ::= b<B>
<S> ::= <empty>
<A> ::= a<S>
<B> ::= bb<S>
```

What is the regular language this RG generates?

Regular Expressions

Regular expressions (RE) are succinct representations of RGs using the following notations.

Sub-Expression	Meaning
x	the single char 'x'
.	any single char except the newline
[abc]	char class consisting of 'a','b', or'c'
[∧abc]	any char except 'a','b','c'
r*	repeat "r" zero or more times
r+	repeat "r" 1 or more times
r?	zero or 1 occurrence of "r"
rs	concatenation of RE "r" and RE "s"
(r)s	"r" is evaluated and concatenated with "s"
r | s	RE "r" or RE "s"
\x	escape sequences for white-spaces and special symbols: \b \n \r \t

Precedence of Regular Expression Operators

The following table gives the order of RE operator precedence from the highest precedence to the lowest precedence.

Function	Operator
parenthesis	`( )`
counters	`* + ? { }`
concatenation	
disjunction	`|`

RE Example 2: Regular Expression Notations

RE	Meaning
abc	the string "abc"
a+b+	$\{a^m b^n : m, n \geq 1\}$
a*b*c	$\{a^m b^n c : m, n \geq 0\}$
a*b*c?	$\{a^m b^n c \text{ or } a^m b^n : m, n \geq 0\}$
xy(abc)+	$\{\text{xy(abc)}^n : n \geq 1\}$
xy[abc]	{xya, xyb, xyc}
xy(a|b)	{xya, xyb}

Questions: What are the following REs?

- foo|bar*
- foo|(bar)*
- (foo|bar)*

RE Example 3: Regular Expressions

- REs are commonly used for pattern matching in editors, word processors, commandline interpreters, etc.
- The REs used for searching texts in Unix (vi, emacs, perl, grep), Microsoft Word v.6+, and Word Perfect are almost identical.
- Examples:
  - identifiers in C++:
  - real numbers:
  - email addresses:
  - white spaces:
  - all C++ source or include files:

Summary on Regular Grammars

- √ There are algorithms to prove if a language is regular.
- √ There are algorithms to prove if a language is context-free too.
- √ English is not RL, nor CFL.
- √ REs are commonly used for text search.
- √ Different applications may extend the standard RE notations.