Regex Speedups

Mark Senn

July 10, 2007

Tools Available

Tools Available

re2c is a tool for writing scanners.

Tools Available

re2c is a tool for writing scanners.
It uses its own (very limited) regular expression syntax.

Tools Available

re2c is a tool for writing scanners.
It uses its own (very limited) regular expression syntax.
How to match strings in middle of input?

Tools Available

re2c is a tool for writing scanners.
It uses its own (very limited) regular expression syntax.
How to match strings in middle of input?

re2xs converts Perl-like (but still limitied) regular expressions to re2c regular expressions.

Tools Available

re2c is a tool for writing scanners.
It uses its own (very limited) regular expression syntax.
How to match strings in middle of input?

re2xs converts Perl-like (but still limitied) regular expressions to re2c regular expressions.
How to match strings in middle of input?

Tools Available

re2c is a tool for writing scanners.
It uses its own (very limited) regular expression syntax.
How to match strings in middle of input?

re2xs converts Perl-like (but still limitied) regular expressions to re2c regular expressions.
How to match strings in middle of input?

Perl 5.10 will contain code to optimize regular expressions.

Tools Available

re2c is a tool for writing scanners.
It uses its own (very limited) regular expression syntax.
How to match strings in middle of input?

re2xs converts Perl-like (but still limitied) regular expressions to re2c regular expressions.
How to match strings in middle of input?

Perl 5.10 will contain code to optimize regular expressions.

Text::Match::FastAlternatives searches input very quickly for a list of strings.

Trie

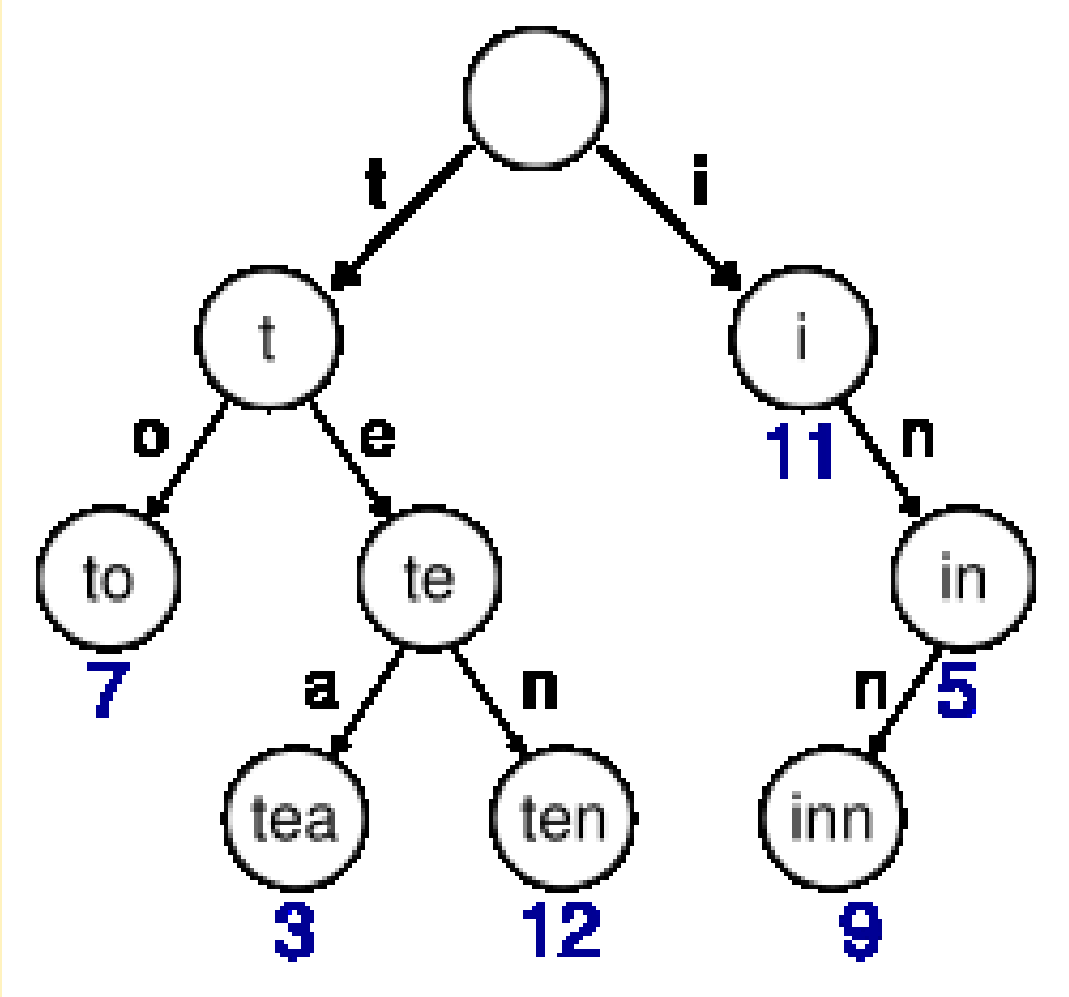
From wikipedia:

> The term **trie** comes from re**trie**val and is pronounced “tree”.

Trie

From wikipedia:

> The term **trie** comes from re**trie**val and is pronounced "tree".



Methods Tested

Methods Tested

Text::Match::FastAlternatives

```
$search = Text::Match::FastAlternatives->new(@string);
foreach (@word)  {
    ($search->match($_))  and  $found++;
}
```

Methods Tested

Text::Match::FastAlternatives

```
$search = Text::Match::FastAlternatives->new(@string);
foreach (@word)  {
    ($search->match($_))  and  $found++;
}
```

regex alternate optimized

```
$regex = join '|', map { quotemeta } @string;
$regex = qr/$regex/o;
foreach (@word)  {  ($_ =~ $regex)  and  $found++;  }
```

Methods Tested

Text::Match::FastAlternatives

```
$search = Text::Match::FastAlternatives->new(@string);
foreach (@word)  {
    ($search->match($_))  and  $found++;
}
```

regex alternate optimized

```
$regex = join '|', map { quotemeta } @string;
$regex = qr/$regex/o;
foreach (@word)  {  ($_ =~ $regex)  and  $found++;  }
```

regex loop optimized

```
foreach (@string)  {  push @regex, qr/$_/o;  }
foreach (@word)  {
    foreach $regex (@regex)  {
        ($_ =~ $regex)  and  $found++,  last;
    }
}
```

Performance

The test program searches all 25,143 words from /usr/dict/words using 1,000 of them as patterns.

Performance

The test program searches all 25,143 words from /usr/dict/words using 1,000 of them as patterns.

The times below are in seconds for one initialization and ten searches through all the words.

Performance

The test program searches all 25,143 words from /usr/dict/words using 1,000 of them as patterns.

The times below are in seconds for one initialization and ten searches through all the words.

Method	Setup	Run	Total
Text::Match::FastAlternatives	0.00	0.84	0.84
regex alternate optimized	0.00	104.81	104.81
regex loop optimized	0.03	190.20	190.23