Theoretical Aspects of Computing – ICTAC 2018

Formalising boost POSIX regular expression matching

Martin Berglund[1], Willem Bester[2(B)], and Brink van der Merwe[2]

[1] Department of Information Science and Centre for AI Research, University of Stellenbosch, Private Bag X1, Matieland, 7602 Stellenbosch, South Africa pmberglund@sun.ac.za
[2] Division of Computer Science, University of Stellenbosch, Private Bag X1, Matieland, 7602 Stellenbosch, South Africa {whkbester,abvdm}@cs.sun.ac.za

Abstract

Whereas Perl-compatible regular expression matchers typically exhibit some variation of leftmost-greedy semantics, those conforming to the posix standard are prescribed leftmost-longest semantics. However, the posix standard leaves some room for interpretation, and Fowler and Kuklewicz have done experimental work to confirm differences between various posix matchers. The Boost library has an interesting take on the posix standard, where it maximises the leftmost match not with respect to subexpressions of the regular expression pattern, but rather, with respect to capturing groups. In our work, we provide the first formalisation of Boost semantics, and we analyse the complexity of regular expression matching when using Boost semantics.