Lightweight Structure in Text

Robert C. Miller

May 2002
CMU-CS-02-134
CMU-HCII-02-103

School of Computer Science
Computer Science Department
Carnegie Mellon University
Pittsburgh, PA

Thesis Committee:
Brad A. Myers, Co-chair
David Garlan, Co-chair
James H. Morris
Brian Kernighan, Princeton University

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy

This research was sponsored by the National Science Foundation under grant no. IRI-9319969 and grant no. IIS-0117658, the Army Research Office (ARO) under National Defense Science and Engineering Grant no. DAAH04-95-1-0552, the Defense Advanced Research Project Agency (DARPA) - US Army under contract no. DAAD1799C0061, DARPA - Navy under contract no. N66001-94-C-6037, N66001-96-C-8506, and the USENIX Association.

The views and conclusions contained herein are those of the author and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of any sponsoring party or the U.S. Government.

Keywords: text processing, structured text, pattern matching, regular expressions, grammars, web automation, repetitive text editing, machine learning, programming-by-demonstration, simultaneous editing, outlier finding, LAPIS

Abstract

Pattern matching is heavily used for searching, filtering, and transforming text, but existing pattern languages offer few opportunities for reuse. *Lightweight structure* is a new approach that solves the reuse problem. Lightweight structure has three parts: a model of text structure as contiguous segments of text, or *regions*; an extensible library of structure abstractions (e.g., HTML elements, Java expressions, or English sentences) that can be implemented by any kind of pattern or parser; and a *region algebra* for composing and reusing structure abstractions. Lightweight structure does for text pattern matching what procedure abstraction does for programming, enabling construction of a reusable library.

Lightweight structure has been implemented in LAPIS, a web browser/text editor that demonstrates several novel techniques:

- *Text constraints* is a new pattern language for composing structure abstractions, based on the region algebra. Text constraint patterns are simple and high-level, and user studies have shown that users can generate and comprehend them.
- *Simultaneous editing* uses multiple selections for repetitive text editing. Multiple selections are inferred from examples given by the user, drawing on the lightweight structure library to make fast, accurate, domain-specific inferences from very few examples. In user studies, simultaneous editing required only 1.26 examples per selection, approaching the 1-example ideal.
- *Outlier finding* draws the user's attention to inconsistent selections or pattern matches — both possible false positives and possible false negatives. When integrated into simultaneous editing and tested in a user study, outlier finding reduced user errors.
- *Unix tools for structured text* extend tools like `grep` and `sort` with lightweight structure, and the *browser shell* integrates a Unix command prompt into a web browser, offering new ways to build pipelines and automate web browsing.

Theoretical contributions include a formal definition of the region algebra, data structures and algorithms for efficient implementation, and a characterization of the classes of languages recognized by algebra expressions.

Lightweight structure enables efficient composition and reuse of structure abstractions defined by various kinds of patterns and parsers, bringing improvements to pattern matching, text processing, web automation, repetitive text editing, inference of patterns from examples, and error detection.

Contents

List of Figures

List of Tables

List of Algorithms

Acknowledgements

This thesis is dedicated to:

- my dear wife Laura, who puts up with a lot;
- my parents Larry and Marian, and my brothers Craig and Eric, for their love and support, and for only rarely asking when I was going to get a real job;
- Dick Kollin and Dave Fraser, whose shameless exploitation of teenage labor gave me my start in this crazy business;
- Charles Leiserson, who encouraged me to go on to grad school at CMU and predicted I'd come back someday;
- Brad Myers, who pointed me in the right direction, gave me the ball, and let me run with it;
- my other committee members David Garlan, Jim Morris, and Brian Kernighan;
- my fellow grad students James Landay, Rich McDaniel, John Pane, Jeff Nichols, and Jake Wobbrock;
- Sharon Burks and Ava Cruse, who made impossible things seem easy;
- Yuzo Fujishima, who kindly provided materials for my user studies;
- the wonderful people with whom I have shared offices (and laughter) over the years at CMU, including George Necula, Fay Chang, José Brustoloni, Adam Berger, Chris Colohan, Hakan Younes, Umair Shah, Bartosz Przydatek, and Sanjit Seshia;
- everyone who gave valuable feedback about LAPIS, including Rich Clingman, Monty Zukowski, Julián Jesús Martínez López, John Padula, Ben Bostwick, Franklin Chen, Chris Long, and John Gersh.