Abstract

A pattern is a word that consists of variables and terminal symbols. The pattern language that is generated by a pattern \( \alpha \) is the set of all terminal words that can be obtained from \( \alpha \) by uniform replacement of variables with terminal words. For example, the pattern \( \alpha = axyax \) (where \( x \) and \( y \) are variables, and the letter \( a \) is a terminal symbol) generates the set of all words that have some word \( aw \) both as prefix and suffix (where these two occurrences of \( aw \) do not overlap).

Due to their simple definition, pattern languages have various connections to a wide range of other areas in theoretical computer science and mathematics. Among these areas are combinatorics on words, logic, and the theory of free semigroups. On the other hand, despite their simple definition, many of the canonical questions in formal language theory are surprisingly difficult for pattern languages.

The present thesis discusses various aspects of the inclusion problem of pattern languages. It can be divide in two parts. The first part (Chapters 3 and 4) directly examines the decidability of the inclusion problem for pattern languages and for a related model.

In Chapter 3, we prove that inclusion for pattern languages remains undecidable even if the number of variables and the number of letters in the terminal alphabets are bounded. Although the proof of undecidability uses a comparatively large number of variables, far fewer variables already allow the simulation of Collatz iterations.

Chapter 4 adapts the proofs of the previous chapter to a superclass of pattern languages, the languages that are generated by regular expressions with variables (or backreferences) (which are based on an extension of regular expression that
can be found in almost every modern implementation of regular expressions). Compared to pattern languages, this language class has greater expressive power, which allows us to extend the proofs of Chapter 3 to undecidability of regularity and inclusion. As a consequence, regular expressions with variables cannot be minimized, and there are non-recursive tradeoffs between regular expressions with and without variables. This holds even if only a single variable is used.

The second part (Chapters 5 to 7) deals with descriptive patterns, the smallest generalizations of arbitrary languages possible within a class of pattern languages ("smallest" with respect to the inclusion relation).

In Chapter 5, we prove that there are languages of which no E-pattern is descriptive. Chapter 6 introduces the concept of descriptive generalization, a learning theoretic model where arbitrary languages are approximated through descriptive patterns. Finally, Chapter 7 examines and disproves a conjecture on a characterization of languages that have no descriptive terminal-free E-pattern.

The thesis is available online at
http://publikationen.ub.uni-frankfurt.de/volltexte/2011/10991

Table of Contents

1 Introduction ........................................... 1
   1.1 On Patterns ....................................... 1
   1.2 On This Thesis .................................... 2
2 Preliminaries ........................................ 5
   2.1 Basic Definitions ................................... 5
   2.2 Patterns and Their Languages ...................... 6
3 Inclusion of Pattern Languages ....................... 9
   3.1 On Inclusion for Pattern Languages ............... 9
   3.2 Definitions and a Preliminary Result ............. 10
   3.3 The Difficulty of Inclusion ....................... 13
   3.4 From Pattern Inclusion to Regular Expressions .. 36
4 Real Regular Expressions: Decidability and Succinctness .... 41
   4.1 On Extended Regular Expressions ................. 41
   4.2 Definitions and Preliminary Results .............. 43
   4.3 Undecidability and Its Consequences ............. 63
5 Existence of Descriptive Patterns .................................. 73
  5.1 On Patterns Descriptive of a Set of Strings .................... 73
  5.2 Preliminaries .................................................... 75
  5.3 Descriptive Patterns and Infinite Chains ....................... 80
  5.4 Existence of Descriptive Patterns .............................. 81
  5.5 Computing Descriptive Patterns .............................. 93
6 Inferring Descriptive Generalizations .............................. 97
  6.1 On Descriptive Generalizations ............................. 97
  6.2 Inferring Descriptive Generalizations ....................... 98
  6.3 Inferring ePAT_{df,\Sigma}-Descriptive Patterns ............... 104
  6.4 Examination of the Class TSL_\Sigma ........................ 111
7 On a Conjecture on ePAT_{df,\Sigma}-Descriptive Patterns .......... 115
  7.1 Technical Preliminaries and Various Conjectures ............. 116
  7.2 Chains, Chain Systems, and Their Languages ............... 118
  7.3 The Languages L^{(k)}_\Sigma .................................. 127
  7.4 The Loughborough Example .................................. 130
  7.5 The Wittenberg Examples .................................. 132
8 Conclusions and Suggestions for Future Research ............... 137

Author’s correspondence address  Dominik D. Freydenberger
Institut für Informatik
Goethe-Universität
Postfach 11 19 32
60054 Frankfurt am Main
Germany