REPORT

on a Thesis for awarding the degree "Doctor of Sciences"

Scientific field: Natural sciences, mathematics and informatics

Professional field: 4.6. Informatics and computer science

Title: Finite-state automata, transducers and bimachines:
algorithmic constructions and implementations

Author: Stoyan Milkov Mihov

Overview

The presented DSc Thesis deals with the theory of the finite-state automata, transducers and bimachines. The main goal of the author is to present constructions of devices with a finite number of states, correctness proofs, as well as working implementations in a language designed by the author to describe such machines. The considered abstract devices have important applications in the processing and storage of natural languages. The dissertation marks a considerable progress towards the definition at a conceptual level of several types of finite-state machines used in speech processing.

This thesis is based on the author's research in the last 15 years. The main problems tackled in this work are the following:

(1) Presentation and self-contained description of the theory of finite automata, transducers and bimachines

(2) Formal proof of soundness of the presented constructions, as well as proof of the equivalence of the various abstract devices.

(3) Development of a language used for description of finite state machines, transducers and bimachines.

(4) Development of implementations together with the corresponding documentations for the presented constructions.

(5) Description of possible applications of the developed techniques.

State of research

My general impression is that the author is well acquainted with the state of the art and the most recent results in the theory of finite state machines. A big part of the considered problems are considered in the field as important theoretically, as well as for the applications. The author demonstrates deep knowledge of his field of research and capacity to apply his knowledge to the solution of important problems.

Methods

In his research the author uses combinatorial and algebraic results, as well as some basic techniques from set theory and the general theory of finite automata.

Brief description of the thesis

This thesis amounts 219 pages of text and consists of an introduction, eight chapters, conclusion and a list of references including 48 items. Below, I give a short description of the topics covered in the eight chapters of this dissertation.

Section 1 contains definitions of basic mathematical notions used in the thesis. These include sets, functions and relations (section 1.1), extensions of functions (section 1.2) languages over arbitrary alphabets (section 1.3), strings and relations over strings (section 1.4). In section 1.5, the monoids are defined as sets with a binary operation, which is associative and has a unit. Further, languages over monoids and operations on monoids are defined.

In section 2, the author introduces finite automata as a generalization of the classical automata. To every word $a_1 \dots a_k$ from the language, recognized by the automaton, a label is $a_1 \circ \dots \circ a_k$ attached, where the expression is computed in the monoid $\mathcal{M}$. The classical automata are obtained as a special case of monoidal automata over the free monoid. Under these conditions, the author proves theorem 2.1.22, saying that every monoidal automaton is the homomorphic image of a classical automaton, and proposition 2.2.1, saying that the family of the languages recognized by monoidal automata is closed under the operations union, monoidal product, monoidal Kleene-star, as well as with respect to monoidal homomorphisms.

In section 2.3, the monoidal regular languages are introduced in analogy with the classical regular languages. The difference is in the requirement that $L_1 \circ L_2$ should be a monoidal language, as well. This definition is more general: the classical regular languages are obtained as a special case of the monoidal languages. Furthermore, the monoidal regular expressions are introduced, and it is proved that the monoidal regular languages are exactly these languages that are recognized by the monoidal finite automata (theorem 2.4.2).

In the last section 2.5 of this chapter, some simplifications of the finite monoidal automata are considered. The first one is the so-called trimming, which consist in deletion of the superfluous states: each state has to be on a path which reaches a finite state. Another operation is the removal of the e-transitions, transitions that have as a label the unit of the monoid. It is proved that for every monoidal finite automaton there exists an equivalent monoidal finite automaton without e-transitions (proposition 2.5.4).

Chapter 3 deals with deterministic finite automata as the most important class of finite automata. In section 3.1 the deterministic finite automata are introduced and several well-known properties of the latter are defined (e.g that every finite language is recognized by a finite automaton). In section 3.2 a construction is described by which given a finite monoidal automaton, one constructs a deterministic finite automaton with a total transition function.

In section 3.3, the author considers some properties of the classical finite automata. The central results her are propositions 3.3.1, 3.3.2, and 3.3.4. In the first one, given a language recognized by an automaton, it is proved that there exists an automaton that recognizes the complementary language. In proposition 3.3.2 a similar statement is proved for the intersection and difference of languages recognized by a finite automaton. Proposition 3.3.4 relates to letter reversal.

Sections 3.4 and 3.5 are devoted to the minimization of finite automata. The notions right invariant relation and compatibility of a relation and a language are introduced. It is proved that if a language is compatible with a right invariant relation of finite index, then the language is recognized by a deterministic finite automaton (proposition 3.4.4). In addition, if $R_{\mathcal{A}}$ is a relation including all pairs of words read on a path from q_0 to the same state, then $R_{\mathcal{A}}$ is right invariant (proposition 3.4.7). Furthermore, for any language L over the alphabet Σ the Myhill-Nerode relation R_L is introduced and several results are proved. The most important of them are the following: theorem 3.4.13, according to which for every classical finite automaton there exists a unique equivalent deterministic finite automaton which is minimal with respect to the number of states; theorem 3.4.14, which says that a language L is recognized by a finite automaton if and only if the index R_L is finite; proposition 3.4.17, which states that an automaton is finite if and only if there exist no different equivalent states. In section 3.5, a construction is presented which given a finite automaton produces an equivalent minimal automaton by consecutively identifying of all equivalent states.

Section 3.6. considers colored deterministic finite automata. These are automata in which all states are "colored" in several colors, i.e. a partition on the final states is given, which, on its part, induces a partition on the recognized words. The main results here are theorems 3.6.12 and 3.6.14. The former states that for every colored deterministic finite automaton there exists a unique equivalent deterministic finite automaton which is minimal with respect to the number of states. The latter theorem states that a deterministic finite automaton is minimal if and only if there exist no different equivalent states.

Finally, in section 3.7, the author considers pseudo-determinization of monoidal finite automata. A monoidal finite automaton is pseudo-deterministic if from every state and by any element of the monoid at most one state can be reached. An automaton is pseudo-minimal if it is a homomorphic image of a minimal automaton. The author proves in proposition 3.7.2 that for every monoidal finite automaton one can construct at most an equivalent pseudo-deterministic finite automaton. Proposition 3.7.4 states that for every monidal finite automaton there exists an equivalent pseudominimal finite automaton.

Chapter 4 is devoted to monoidal multi-tape finite automata and finite-state transducers. In section 4.1, monoidal multi-tape automata are introduced. In such automata the input alphabet is a direct product of n monoids. So, the language recognized by an n-tape automaton

is an n-ary relation. The monoidal multi-tape automata have some expected properties that are described in proposition 4.2.1. This implies that the class of the finite monoidal multi-tape relations is closed with respect to the cartesian product, projections and inversion. In section 4.3, the classical finite n-tape automata are defined (as n-tape monoidal automata over the free monoid), along with the finite n-tape letter automata. In sections 4.4 and 4.5 the author introduces monoidal finite transducers and classical finite transducers, respectively, as 2-tape finite automata over an alphabet of the form $\mathcal{M}_1 \times \mathcal{M}_2$, where $\mathcal{M}_1$ is a free monoid, respectively $\mathcal{M}_1$ and $\mathcal{M}_2$ are free monoids. Finally, in section 4.6, several results are proved, that lead to a procedure for deciding the functionality of a classical finite transducer. All this is implemented in section 8.

In chapter 5, the author investigates finite transducers. In Section 5.1, the notions of a monoidal subsequential transducers and a classical subsequential transucers are introduced. The important notions of sequential distance between two words and a function of bounded variation are introduced. The latter is a function for which close words are transformed again to close words. The next goal is to investigate classical finite transducers of bounded variation. In section 5.2 the author proves that every functional classical transducer of bounded variation can be transformed to an equivalent classical subsequential transducer. Further, in section 5.6, it is shown how these results can be generalized to monoids, different from the free monoid (e.g. for the monoid of the natural numbers with respect to addition).

In section 5.4, the author defines a relation, similar to the Myhill-Nerode relation, by which he constructs a minimal subsequential transducer. In section 5.5, a procedure for minimization of subsequential finite transducers is presented.

In chapter 6, the author investigates bimachines, described as finite computing devices which consist of monoid alphabet, two deterministic finite automata with a common alphabet and a partial function, called the output function.

In section 6.2, the connection between the regular functions on words and the bimachines is investigated. It is proved that for every monoidal bimachine $\mathcal{B}$ there exists a monoidal finite transducer $\mathcal{A}$, for which $O_{\mathcal{B}} = L(\mathcal{A})$ (Proposition 6.2.1). Conversely, for every trimmed functional transducer with output the monoid $\mathcal{M}$ there exists a monoidal bimaxine $\mathcal{B}$, for which $L(\mathcal{T}) = O_{\mathcal{B}}$. In section 6.3, the notion of pseudominimization of a monoidal bimachine is introduced. In section 6.4, the so-called direct composition of classical bimachines is investigated.

Chapter 7 contains a description of the language $C(M)$ which is used in last chapter for description of the algorithm presented in the dissertation. A compiler is developed which compiles a program to code in C. In the last chapter 8, implementations are presented for algorithms for different abstract nmachines: finite automata, classical finite transducers, deterministic finite transducers and bimachines.

Main results in the thesis

The main results in this DSc Thesis are the following:

(1) The theoretical foundations of finite automata, transducers and bimachines are presented.

Proof for the correctness of the presented constructions, as well as for the connections between them, are given.

(2) A method for testing the restricted variation of a classical finite transducer is presented.

(3) A programming language $C(M)$ for realization of the algorithms in the dissertation is presented. A compiler is created which compiles $C(M)$-programs to C-code.

(4) Implementations of algorithms for different abstract machines are presented: finite automata, classical finite transducers, deterministic finite transducers and bimachines.

(5) Applications of the developed theory to important practical problems are given.

Remarks and comments

I have the following remarks, questions and comments related to this thesis:

(1) The authors summary is rather scarce. As things are, it is just a collection of definitions and results. At places some explanations are needed to make the text intelligible without reference to the main text of the thesis. For instance in the definition of sequential distance the symbol $u \wedge v$ is not defined; in proposition 3.3.3, it is not explained what ρ is, and so on.

(2) Some of the notions are not correctly translated in Bulgarian. (refinement)

(3) What is the difference between a proposition and a theorem? Both are used in the text.

Publications related to the thesis

This thesis follows closely the book by the author and K. Schulz entitled "Finite-State Techniques: Automata, Transducers and Bimachines", published by Cambridge University Press. The results of this thesis are published in 12 papers and one chapter from a book. Three of the papers are in a journal with IF:

- Computational Linguistics (Q1:2000,2004) – 1.657

- Theoretical Computer Science (Q3: 2011) – 0.667

Seven of the remaining papers are in journals and conference proceedings with SJR: International J. of Document Analysis and Recognition, ACM Transactions on Speech and Language Processing, Lecture Notes in Computer Science etc. A chapter from the book Finite State Techniques: Automata, Transducers and Bimachines from the sequence Cambridge Tracts in Theoretical Computer Science, Cambridge University Press is also used. The presented publications meet the minimal national requirements as given in the corresponding documents.

Authorship of the obtained results

In six of his papers, Stoyan Mihov has one co-author, five papers are with two co-authors, and one with three co-authors. A letter from one of the co-authors is resented explaining that the contribution of Mihov in the joint book is significant. It is stated that the research concerning the language $C(M)$, as well as the implementations of finite-state machines are entirely his own work. In the remaining joint publications the contribution of the author of this thesis is also significant.

Citations

The candidate has attached a list of 227 citations of the papers used in this thesis. It is beyond any doubt that the results of Mihov are well-known and highly valued in his professional community.

Authors summary

The authors summary are made according to the regulations and reflect properly the main results and contributions of this thesis.

Conclusion

This thesis is focused on problems from the theory of the finite-state machines that are of great importance for the applications, especially those related to language processing, language storage, and phonetization. The author develops new algorithms for construction and assessment of finite automata, which marks a considerable progress in an important field of informatics.

I am deeply convinced that the presented thesis "Finite-state automata, transducers and bimachines: algorithmic constructions and implementations" by Stoyan Milkov Mihov contains results that are an original contribution to the theory of the finite state machines. The candidate demonstrates deep knowledge of the theory and capacity to develop it in new and important way. With this, he meets the legal national requirements prescribed by the law plus the specific ones of IICT-BAS for the professional field 4.6 Informatics. I assess **positively** the presented DSc Thesis and recommend this panel to award **Stoyan Milkov Mihov** the scientific degree "Doctor of Sciences" in the scientific field 4. Natural sciences, mathematics and informatics Professional field 4.6 "Informatics".

Sofia, 28.03.2020 g. Member of the Scientific Panel:

(Prof. Ivan Landjev)