Referee's report on the PhD Thesis

Referee: Prof. RNDr. Alexander Meduna, CSc
PhD Thesis: On Indexes of Ordered Trees for Subtrees and Tree Patterns and Their Space Complexities
Author: ing. Martin Poliak
Supervisor: Doc. ing. Jan Janousek, PhD

Summary

The thesis deals with using automata for the problem of indexing trees for subtrees and patterns. Such an indexing automaton accepts all subtrees or all tree patterns that match a given tree. Given a tree with n nodes, there is just n subtrees of the tree and there can be an exponential number of different tree patterns that match the tree.

Organization

The thesis has standard structure and consists of seven chapters. The first two chapters contain Introduction and Basic notions. The third chapter describes related work. The other three chapters are devoted to the three main results, as they are described above. The last section contains conclusion.

Results

The thesis contains three main results of the author. First, a deterministic pushdown automaton, called tree compression automaton, is presented. This pushdown automaton accepts all subtrees of a given tree in a linear notation and can be used for computing subtree repeats. Furthermore, the tree compression automaton can be used for compressing trees. Second, a new full index of a tree for tree patterns is described. That index uses also a well-known string compact suffix automaton. The presented index has linear size to n. The acceptance of an input tree pattern of size m is performed in time linear to m plus the number of occurrences of the pattern in the tree. Third, the thesis contains a detailed discussion on space complexities of deterministic finite tree automata and deterministic pushdown automata that accept all tree patterns matching a given tree. Various structures of trees are studied and the corresponding sizes of the indexing automata are shown and proved.

The thesis presents new original results of the author. Two new particular indexing algorithms and their properties are introduced and studied. Both these indexes are based on automata. The third part is focused on space complexities of indexes by finite tree and pushdown automata. This part contains new theoretical results.

The thesis represents a significant contribution to the topic of processing trees with the use of automata. All the proofs in the thesis are thoroughly written. The quality of the text shows a high level of knowledge of the author.
Disadvantages

In a cooperation with his supervisor, the author should improve his publications. He should publish more international-level papers based on this work.

Questions

The problem of indexing trees for tree patterns effectively is still an open problem area. Both particular indexing algorithms and studies on complexities can be investigated. The algorithms from the thesis can serve as the basis thesis for future research. It would be interesting to consider also indexing for approximate patterns instead of exact patterns. Also “oracle” variants, inspired by factor oracle automaton, of the indexing automata can be investigated for a memory efficient implementations.

Consider the open problem area sketched above. Are there of some interest? Have you studied any of them? If not, do you plan to study them?

Conclusion

The PhD thesis has achieved the defined objectives, used appropriate methods and presented new knowledge, which is both significant in theory and applicable in practice. Therefore, this thesis satisfies all the Ph.D. requirements.

The thesis is ready for its final defense.

Prof. RNDr. Alexander Meduna, CSc.
Department of Information Systems
Faculty of Information Technology
Brno University of Technology

January 28, 2018
Hamilton, December 25, 2017

Review of Ph.D. thesis titled *On Indexes of Ordered Trees for Subtrees and Tree Patterns and Their Space Complexities* by Ing. Martin Poliak, submitted to the Faculty of Information Technology, Czech Technical University in Prague, in partial fulfilment of the requirements for the degree of Doctor.

I reviewed the thesis and rate it as *excellent*, and thus, my recommendation is to proceed with the defense. In the following I will justify my recommendation.

I reviewed the thesis in three distinct points of view: (1) the quality and comprehensibility of the presentation of the thesis itself, (2) the quality and originality of the research presented in the thesis, and (3) the level of contribution to the body of knowledge in the chosen area.

(1) The quality and comprehensibility of the presentation of the thesis.

The thesis comprises 7 chapters and a bibliography. The first chapter provides a motivation for and a brief introduction to the chosen research area and includes a clearly formulated Problem Statement. Chapter 2 provides a concise presentation of the basic terms and notions required for understating of the presented research. Chapter 3 presents a succinct overview of related research, very well structured from the pedagogical point of view: first the (simpler) string indexing is presented, and then the more complex tree indexing is presented; without first presenting the string indexing, the material would be hard-to-access by a more general audience of computer scientists and software engineers. Section 3.5 presents XLM indexing methods, which, in my opinion is neither necessary nor helpful for the main research aspects of the thesis (I recommend leaving it out). Chapter 4 presents the formalism and the results concerning the deterministic Tree Compression Automaton. Ing. Poliak managed this chapter in an excellent manner walking the fine line between insufficient and excessive formalism. This chapter has a textbook quality in the presentation and phraseology. Chapter 5 presents a linear index for trees to access tree patterns, and again the chapter is very well structured and presented. Chapter 6 discusses the space (storage) requirements of the various pushdown automata. Chapter 7 provides a concise conclusion with a clear delineation of the author’s contribution and some suggestions for further research.

The whole thesis is very carefully written; I found just a handful of insignificant typos. The level of English is extremely good; of course there were a (very) few awkward phrases or formulations, but none of them was detrimental to presentation. The overall look of the thesis is very professional and indicates quite a careful effort to help the reader, which I noted with a high regard. My only slightly negative comment concerns the presentation of the bibliography – it should be unified, either all authors should be with full first names, or all should be with just initials (cf. [1] Alfred V Aho ... and [5] P. Bille ...) . Whichever style is fine. The references containing URL’s should be updated (though it is a futile task), but at least...
they should be up-to-date at the time of the defense. For instance, the URL in the reference [3] is already obsolete. In several places (I can provide a detailed localization if need be) I encountered expressions $O(2n)$ or $O(3n)$ etc. This is slightly confusing as $O(n) \equiv O(2n)$. Since it occurs several times, the author had a reason for it. Either it should be explained why this unusual notation is occasionally used, or some different notation expressing what the author intended to say should be employed. Overall evaluation of the quality and comprehensibility of the presentation of the thesis is excellent.

(2) The quality and originality of the presented research.
The research aspects are presented in Chapters 4, 5, and 6. The main area is what the author calls arbology, or more commonly called “Pattern Matching on Trees”. The research presented here is a part of an effort directed and motivated by the Arbology Project at the Technical University and seamlessly extends the previous research of some of the former Ph.D. students and professors of the department. The research is focused on the automata theory approach to the pattern matching on trees and is of the highest standard. Both the topic and its research resolution are highly original, as is the whole Arbology project. A significant part of the thesis research (see Chapter 6) is a major part of the paper On Space Requirements of Indexes of a Tree for Tree Patterns Based on Deterministic Pushdown Automata submitted to the journal Theoretical Computer Science and Tree Compression Pushdown Automaton that appeared in the journal Kybernetika in 2012 and conference contribution Full and Linear Index of a Tree for Tree Patterns.
The theorems and results presented are correct. I was a bit confused with the proof of Theorem 4.14 (page 35) with the second type of the transition rules – what is $q'$? The formulation of Theorem 4.23 is quite confusing: it seems to me that the author means to say “can have at most $x$” rather than “can have only $x$” as the latter formulation really means to “have exactly $x$”. Than the conclusion of the theorem says “cannot have less than”. It should be reformulated to indicate explicitly which part describes the upper bound and which describes the lower bound. In Definition 6.7 (page 77) it is not clear what $A_n$ refers to. It is clear, that the research presented and discussed in the thesis is of a high quality as required now for international recognition, and so my evolution of this aspect was excellent.

(3) The level of contribution to the body of knowledge in the chosen area.
The contribution of the author is threefold and contained in the research Chapters 4, 5, and 6. It corresponds to the topics of the chapters. In Chapter 4, a deterministic pushdown automaton for tree compression is presented. It represents a significant result in automata theory. In Chapter 5, a linear index for a given tree for tree patterns is presented, again, a significant result. I would appreciate a bit more elaboration on the meaning of the term “tree pattern”. Finally, in Chapter 6, the space complexities of deterministic pushdown automata are discussed. I particularly enjoyed that chapter and found it refreshing as most researchers in automata theory do not pay that much attention to space requirements. The methods
employed and the level of presentation significantly contribute to the body of knowledge. Overall, I rated this aspect of the thesis also **excellent**.

RNDr. Frantisek (Franya) Franek, Ph.D., L.E.L.
Professor
Department of Computing & Software
McMaster University, Ontario, Canada L8S 4K1
e-mail:  franek@mcmaster.ca
http://www.cas.mcmaster.ca/~franek/
Review of PhD thesis

On Indexes of Ordered Trees for Subtrees and Tree Patterns and Their Space Complexities

Author: Martin Poliak

The thesis has on three parts concerning indexing of trees for subtrees and for tree patterns: tree compression automaton, a full and linear index of a tree for tree patterns based on a tree pattern pushdown automaton (PDA), and space requirements on indexes based on finite tree automaton (FTA) and tree pattern PDA. The research builds on the work of the Arbology research group at the Faculty of Information Technology CTU in Prague.

The introductory Chapter 1 describes the problem statement and structure of the thesis. Chapter 2 contains all necessary definition, like trees, tree patterns, PDA, FTA, subtree PDA, tree pattern PDA including associated construction algorithms. As linearized trees are considered, important notions like suffix trie and suffix tree are defined in Chapter 3 including related works concerning tree indexing and tree compression.

The core chapters 4-6 present in detail three own author's contributions mentioned above. Chapter 4 explains Tree Compression Automaton (TCA) and the algorithms for its construction for set of ordered trees \( T \), i.e. TCA(T). The algorithm constructs even a general TCA tree, accepting all subtrees of trees from \( T \). The author also discusses the size of the constructed automaton. The associated decompression algorithm is able to reconstruct a subtree \( t \) of a tree from \( T \) entering index of \( t \). This is appropriate to search of patterns in the compressed tree. TCA can be used as an index of a tree, subtree matching, and searching subtree repeats in tree \( t \). The algorithms are proved as theorems, their time and space complexities are discussed in detail. Experiments comparing TCA and a grammar-based algorithm BPLEX have shown that TCA mostly achieves better compression ratios.

Chapter 5 is devoted to construction of index for a single subject tree \( t \) and to searching tree pattern occurrences in \( t \). Time and space complexities are proved. Chapter 6 considers indexes of a tree for tree patterns based on FTA and on tree pattern PDA. Chapter 7 summarizes the work and offers some suggestions for further research.

The application domain for trees considered in this thesis is XML data structures. This choice seems appropriate, since XML data is still widely used in practice. Its compression and indexing are still up to date theme.

Text comments:

p. 26 - it not exactly true that XML documents store data in a hierarchical rooted tree structures. It holds when we do not consider attributes ID, IDREF, and IDREFS. Also a remark stating that the XML trees are unordered is not true. For example, the order between siblings is significant for XPath axes preceding-sibling and following-sibling.

Style: The text is well structured, it contains meaningful definitions and explanatory examples and figures. The structure of particular chapters is not always compatible. For example, some chapters have explicit conclusion, some not. The thesis is clearly written, the English is very good.

The thesis demonstrates that the author achieved valuable results in the specified field. They were evaluated positively by a research community, published and presented on international conferences. In accordance with par. 47, letter (4) of the Law Nr. 111/1998 (The Higher Education Act) I recommend the thesis for presentation and defense with the aim of receiving the PhD degree.

Prague, December 8, 2017

Prof. RNDr. Jaroslav Pokorný, CSc.