Report on the Doctoral Thesis: (Nonlinear) Tree Pattern Indexing and Backward Matching

Author: Ing. Jan Trávníček

The dissertation thesis of Jan Trávníček is devoted to the problem of finding all occurrences of tree patterns in a given tree. Trees are widely used in applications (e.g. data structures), and therefore they constitute an area which is in the center of research for long time. In the thesis, two main approaches to the problem of finding all occurrences of tree patterns are studied. The first is based on preprocessing and forming a complete index of the subject tree from which all occurrences of (nonlinear) pattern can be discovered. The second approach results in a tree pattern matching algorithm designed for different types of tree representations (prefix and postfix ones).

The thesis is divided into six chapters, and the main new results are contained in Chapters 4 and 5. The first three chapters contain and introduction, theoretical background (Chapter 2), and previous results and related work (Chapter 3).

Chapter 4 deals with the problem of indexing. It contains a presentation of a new tree pattern pushdown automaton and a nonlinear tree pattern pushdown automaton. The latter representing a complete index of a given tree for nonlinear patterns. Algorithms that are given here can be easily generalized for dealing with unranked trees.

Chapter 5 studies the second approach to the problem of finding all occurrences of a tree pattern in a given tree. It brings backward linearized tree pattern matching algorithms that can deal with different linear representations of trees (different types of prefix or postfix representations). The algorithms are also generalized to work with nonlinear tree patterns.

For all algorithms presented in the thesis, their time and space complexities are investigated and their upper bounds are calculated and proved. The algorithms given in the thesis were implemented and thorough comparison with regard to the time requirements was conducted and documented.

The methods used in Chapter 5 are based on methods known in stringology but they had to be adopted to more complicated research area. To my opinion all methods used in the thesis are appropriate.

Chapter 4 contains mainly results that were already published, either in conference papers or in a journal publication. In Chapter 5, the author included results/algorithms that have not been published yet.

The thesis is well structured and it is written in satisfactory English. The goals of the thesis were fulfilled. The author presents his ideas clearly. Each algorithm is illustrated on at least one example. The thesis ends with summary and suggestions for the future research work (Chapter 6). The bibliography is broad, thorough and

current. I appreciate that the author explicitly stated which parts of the text that contain results from published papers constitute his own work.

However, the thesis is not void of some mistakes (and typos). Let me mentioned two of them:

- Definition of postfix bar notation is not correct (it is a definition of prefix bar notation), page 9.
- Definition 4.1.1. in the text is not a definition.

Despite of minor mistakes, the thesis is of a high scientific quality and it brings new important results. The main part of the results presented here were already published. To conclude, the author proved the ability to conduct independent research and achieve scientific results. In accordance with par. 47, letter (4) of the Law Nr. 111/1998 (The Higher Education Act) I recommend the thesis for the presentation and defense with the aim of receiving the Ph.D. degree.

Prague, 6. 3. 2019

Prof. RNDr. Marie Demlova, CSc.
Department of Mathematics,
Faculty of Electrical Engineering, CTU

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
FAKULTA INFORMAČNÍCH TECHNOLOGIÍ

Ph.D Thesis Review

Name of the student: Ing. Jan Trávníček

Title of the thesis: (Nonlinear) Tree Pattern Matching

Reviewer: doc. Mgr. Adam Rogalewicz, Ph.D.

Institution: Faculty of Information Technology, Brno University of Technology

Overview of the Results

The thesis tackle the problem of tree pattern matching. The problem of linear pattern matching can be stated as follows: Given a tree and a tree pattern (a tree with possible special symbols S in leaves representing any subtree), find all occurrences of the pattern in the tree. The nonlinear variant is a generalization, where the pattern can also contain special variables in the leaves, where each variable represent a single subtree, which must be present at all positions of the variable. The techniques proposed in the thesis are generalization (but not straightforward) of the techniques for string matching, where the tree structure is taken into an account. The thesis is divided into two main parts representing two main approaches: (1) Tree indexing and (2) direct pattern matching.

The tree indexing option is suitable in the cases, where a set of queries is performed on a single tree. First, an indexing structure for the given tree is created, and then the particular queries can be quickly answered. The author propose several different approaches targeting mostly nonlinear tree patterns. At the end of the section, an experimental evaluation on a huge set of examples of all the mentioned methods (the proposed ones as well as the state-of-the-art ones) is performed.

The direct pattern matching is an option in the cases, where there is going to be just a few queries on a given tree. The search is performed on-the-fly and a lot of optimizations (like skipping parts of the tree where no match is surely possible) is performed in order to obtain a sublinear time complexity. The thesis propose several variants of the matching algorithm, which are principally close each to other. An experimental evaluation is presented at the end of the section, with a minimal performance difference between variants of the proposed method.

Formal Structure and Organization of the Dissertation

The thesis is written in English. The writing stile is good and there are no problems with understanding. The structure of the thesis is also good. First, the basic notions are defined (chapter 2), then the state-of-the-art methods for string and tree matching are presented (chapter 3), and finally, original results of the author for tree indexing (chapter 4) and direct pattern matching (chapter 5) are presented.

Completion of the Dissertation Objectives

The objectives of the dissertation were to transfer all missing techniques used in string indexing and matching into the area of tree indexing and matching. The objectives were fulfilled.

Assessment of the Methods used in the Dissertation

This dissertation consists of a basic research in the *"Arbology"* research area co-founded by doc. Jan Janoušek, the supervisor of the author. The author systematically transfer all missing techniques from string indexing and matching to tree indexing and matching and evaluate all its results. The translation of particular algorithms from strings to trees is not a straightforward option due to fact that trees compared to strings has nontrivial structure. Therefore all the proposed algorithms are nontrivial extension of the string ones.

The dissertation is based on 3 conference papers and 1 journal paper (in the journal with IF), where the core of the thesis was published. This is, according to my opinion, sufficient for the Ph.D. thesis. Moreover, the author has 3 more conference publications in the areas related to this thesis.

Remarks for the Thesis

- pg. 9: post_bar(t) should be $\wedge post\_bar(b1) \ldots post\_bar(bn)\ a$
- pg. 14, bottom: DRTFA called inconsistently "top-down TA"
- pg. 15: double negation in *Unreachable states are not reachable by no sequence ...*
- pg. 20, suffix array: It would be great to add an example how the suffix array works.
- sec. 3.1: It is not clear whether you can always determinize the subtree pushdown automaton.
- sec. 3.2: Can you always determinize tree-pattern PDA?
- sec. 4.2: tnsl(q) is defined via sequences of characters x, but example 4.2.2 is simplified to x be a single character.

- Algorithms 12 and 13 are not mentioned in the text
- pg 46, example 4.2.7: Author mentioned standard deteminization. But no link to literature is provided. This is related to my comment for sections 3.1 and 3.2.
- pg 110, table 5.12: There is a redundant a2 in 6th row.
- Algorithm 37, line 10: Bug. Correct should be *if pref_ranked_bar(pattern)[j]* $\in \{S\} \cup \chi$
- Algorithm 37, line 16: Bug. Correct should be *if variables[pref_ranked_bar(pattern)[j]]* ...
- pg. 115: Does the times contains the construction of SRT_pref_bar?
- section 5.9, second paragraph: The claim in this paragraph is inconclusive. The difference between the variants is within a measurement error.

Questions for the Defense

- Why can you allways determinize the PDAs described within sections 3.1, 3.2 and 4.2?
- What is the complexity of Algorithm 37, when you include construction of SRT_pref_bar? What about comparison of Algorithm 37 (including the time for construction of SRT_pref_bar) with a possibility to create indexing structure and then perform a query (as described in chapter 4)?

Conclusion

The author of the dissertation proved the ability to conduct research and achieve scientific results. In accordance with par. 47, letter (4) of the Law Nr. 111/1998 (The Higher Education Act) **I do recommend** the thesis for presentation and defense with the aim of receiving the Ph.D. degree.

Brno, 22nd February 2019

Adam Rogalewicz

FIT BUT

Dr Johanna Björklund
Umeå University
90187 Umeå, Sweden
+4670 603 94 59

29 December, 2018

Doc. RNDr. Ing. Marcel Jiřina, Ph.D., Dean
Office of Science and Research
Faculty of Information Technology CTU in Prague
Thákurova 9, 160 00 Praha 6

Dear doc. Jiřina and members of the dissertation committee,

What follows is a review of the dissertation by Jan Travnicek submitted to the Faculty of Information Technology CTU in Prague Technical for the degree of Doctor. The dissertation is titled *(Nonlinear) Tree Pattern Indexing and Backward Matching* and comprises 133 pages, divided over 6 main chapters together with an appendix that holds previously omitted proofs. The content is distributed accordingly:

Chapter 1, *Introduction*, introduces the area, motivates the work, and summarizes results.

Chapter 2, *Theoretical background*, provides preliminary knowledge on finite-state string and tree automata, and on different linearisation schemes.

Chapter 3, *Previous and related work*, summaries earlier efforts on pattern matching.

Chapter 4, *Main results in tree indexing*, introduces nonlinear tree pattern pushdown automata and new methods for indexing trees for querying by linear and non-linear tree patterns.

Chapter 5, *Main results in tree pattern matching*, presents different methods for computing backward tree pattern matching, using various forms of linearizations and combinations of linear and non-linear patterns.

Chapter 6, *Conclusion*, summarizes the results on tree indexing an nonlinear pattern matching, and outlines directions for future work.

Up-to-dateness of the dissertation

The sections treating the theoretical background and related work give a fair overview of the state-of-the art. Pattern matching is a topic of perennial interest, due to its fundamental nature and broad applicability.

Formal structure and organization of the dissertation

The thesis has a simple and coherent structure, and the line of argumentation is easy to follow. There are however some redundancies in the presentation that could be omitted to make it even clearer, for example, the division of the pattern matching approached into

those that preprocess the pattern and those that preprocess the subject tree, is repeated several times in the introduction. This is not a big concern, but something to bear in mind for the future.

Completion of the dissertation objectives

The thesis considers the pattern-matching problem for trees. Two varieties are considered: In the first, the subject tree is preprocessed to allow the efficient matching of patterns. In the second, the pattern is fixed, and preprocessed to allow efficient search in many subject trees. Also the patterns are taken to be of two types. The first is linear patterns that intuitively consists of trees with gaps in them, that can be filled with any suitable subtree. The other is nonlinear patterns, where some gaps need to be filled by the same subtree. The thesis deals with all of the above variations, and considers solution approaches based on various type s automata and push-down techniques. The objectives are thus satisfactorily met.

Assessment of the methods used in the dissertation

The thesis uses a combination of deductive reasoning and empirical research, which makes the results very convincing. In my opinion, many of the proofs could have been more formal and used, for example, the establishment of invariants or induction on the index structure to reach their aims. All in all, however, the methodology seems sound and appropriate for the task.

Evaluation of the results and contributions of the dissertation

The thesis contributes to the field of pattern matching for trees. Although several of the data structures and algorithms are already known for strings, the situation is more complex for trees, and the generalisation to this domain is clearly non-trivial.

The thesis has several strengths that are worth noticing. One is the generous use of examples and illustrations. Another is how complex algorithms are presented as the composition of simpler ones. It was also a good idea to separate out some of the less interesting proofs in an appendix, at it makes the body of the thesis more interesting. All-in-all, the thesis represents a solid and useful body of work.

General comments

What follows is a list of more or less general remarks, that can be used to improve the thesis, or as advice for future work. (That is, do not feel forced to include the changes if time is short).

Avoid repetitions in the text, e.g., the second clause of the first sentence of Chapter 1.1 and the first sentence of Chapter 1.3 are almost identical.

p.3 Add a reference to the claim that the backward string matching algorithms often perform sublinearly in practice.

Take care to introduce new notions the first time you use them, such as "bad character shift". I discovered that this term is later explained in full, so perhaps you could move that explanation to an earlier point in the work?

The theoretical background is not very formal:

- Please give the domains from which elements are taken.
- The definition of connected does not work for directed graphs, because according to it, a tree in which all edges point away from the root would not be considered connected.
- Give a formal definition of unranked tree. It is not simply a tree that is not ranked, because every ranked tree is also an unranked tree.
- Try to avoid starting a new paragraph on every line as you do on the first pages of the background.
- The definition of post-bar seems to be the same as that of pre-bar.
- In the last line before Example 2.3.1 it would be better to give the domain and co-domain of the function rather than example elements
- The definition of position heap, dual heap, and suffix array are not clear. Either make them so detailed that the constructions can be followed without consulting the references, or give them in more intuitive form, and avoid using concepts not explained, such as maximal reach pointers. Examples do not help unless you have at least given a high-level definition of the construction.

p.27 In your definitions, make it clear whether you define a construction or a property. For example, Definition 3.2.1 describes a property of the automaton, not a particular type of automata.

Lemma 3.2.7 is not satisfactory, as it does not account for the size of the alphabet. It would be more valuable to know what the corresponding result is when the alphabet is a parameter in the computation.

Add references to all Lemmas and Theorems in chapter 3 that are not the result of your own work.

In Definition 4.2.1 (and also in general): If you have stated that $S, X, Y, \ldots$ are not in the alphabet, then you don't have to write $A \setminus \{S, X\}$.

In Lemma 4.2.8, and in general, try to formulate complexity results in easily measurable properties of the inputs, e.g., their sizes or heights. If you need something more complex, such as the number of distinct prefixes, then give an upper bound of their number based on more standard parameters so that the result become comparable to other people's work.

When writing cursive text in math mode, use the latex command mathit to improve kerning (i.e., letter spacing).

On p.115, add references to the baseline algorithms that you use, if these are described in literature.

Please have a look at *Simulation relations for pattern matching in directed graphs* by Björklund and Öhman, published in Theoretical Computer Science 2013. As it considers pattern matching in directed graphs (which is a generalisation of trees) and preprocesses the subject graph to reduce matching times, some of your methods might carry over to this domain.

Detailed remarks

p.iii "adapted to handle" -> "be adapted to handle"

p.iv Please motivate the stated number of linear and nonlinear patterns at the top of the page. Also, you repeat these figures in the Introduction, and I think once is enough

p.iv "it therefore and" -> "it therefore"

p.xxii There is a spurious "*" after the list of algorithms

p.1 Add reference to the listed application areas, wherever possible

p.2 "are most widely used" -> "are the most widely used"

p.2 "in many methods" -> "in many areas"

p.4 "a new and first" -> "a new"

p.5 "thesisbuilds" -> "thesis builds"

p.10 Do not capitalize characters after semicolon, such as is done in the first line of Example 2.2.4

p.11 "another special wildcard symbols" -> "other special wildcard symbols"

p.12 You define x^* twice

p.25 "however that are necessary" -> "however they are necessary"

p.29 The second transition on the 5th line in the proof of Theorem 3.2.4 seems malformed.

p.30 The sentence starting "The deterministic tree pattern PDA" is not a complete sentence.

p.35 "and later similar algorithm" -> "and later a similar algorithm"

p.35 "hence node labelled by" -> "hence a node labelled by" (in general, try to make sure that you don't forget articles)

p.51 "tales" -> "tails"

p.55, 80, 104, 106 "can't" -> "cannot"

p.56 Add indices to the summation signs in Lemma 4.4.2 and its proof

p.64 In the last line of the proof of Lemma 4.5.20, I think you should use ordo notation since the bound $7n$ is probably not exact.

p.68 In the proof of Theorem 4.6.7, don't use the cardinality of the infinite set $\{X, Y, \ldots\}$. Instead, define the subset of these variables that actually occur in the pattern and take the cardinality of that.

p.79 "and notations defined further in the thesis" -> "and notations defined further on"

p.88 There is some redundancy in the description at the top of the page:

"Lengths of shifts strongly depend on the position of the symbol S in the pattern. Shifts are longer with increasing the distance of the symbol S from the end of the pattern."

"The lengths of the shifts depend on the position of the last wildcard S in pattern p – the closer to the end of the pattern the last occurrence of symbol S is, the longer shifts are performed."

p.92 "symmetricity" -> "symmetry"

p.102 "Meaning, the maximal length" -> "This means that the maximal length"

p.116 "linearised backward linearised tree pattern matching"? (two 'linearised')

Questions for the defense

In the experiments, how many variables, and approx. how many occurrences of each variables did you use in the empirical experiments?

Can the results be translated to unranked trees by using a binary encoding?

Would it be useful to add additional stacks to the pda to remember repeated subtrees?

Overall evaluation

Based on the above considerations, I find that Jan Travnicek has proved the ability to conduct research and achieve scientific results. In accordance with par. 47, letter (4) of the Law Nr. 111/1998 (The Higher Education Act) **I recommend the thesis for the presentation and defense** with the aim of receiving the Ph.D. degree.

Yours faithfully,

J. Björklund

Dr. Johanna Björklund
Assoc. Prof., Dept. Computing Science
Umeå University, Sweden