BRNO UNIVERSITY OF TECHNOLOGY

FACULTY OF INFORMATION TECHNOLOGY

Referee's report on the PhD Thesis by Ondřej Guth

Referee: Prof. RNDr. Alexander Meduna, CSc
Date: May 20, 2014

PhD Thesis: **Searching Regularities in Strings using Finite Automata**

Author: Ing. Ondřej Guth
Supervisor: Prof. Ing. Bořivoj Melichar, DrSc

Summary

This dissertation thesis deals with particular types of string regularities, covers and seeds. A factor of a string is its cover, if the string may be constructed using just superpositions of the cover. Seed of a string is a generalization of cover such that seed is a cover of an extension of that string. In this dissertation thesis we focused on algorithms for computation of all k-approximate covers or seeds of a string. As a measure of approximation within the set maximum distance k, the Hamming distance is used. The main contribution of this thesis are algorithms for computing all k-approximate covers or seeds of a string based on formalism of finite automata.

Covers, and their generalization, seeds, are kinds of repetitive structures of a string and examples of typical regularities of strings. Searching regularities in strings is an important problem in many areas of science. In molecular biology, repetitive elements in chromosomes determine the likelihood of certain diseases. In computer science, repetitive elements in strings are important in data compression, speech recognition, coding etc. String regularities may be seen as a description of repetitive structure of a string. Apart from covers and seeds, one example is period. Every string w may be written as $w = p^i s$, where s is a prefix of p and the length of p is called period. Although period describes structure of a string in an interesting way, for many strings, it is too strict and $|p| = |w|$. Therefore, the notion of quasiperiod and even less strict notion, seed, have been introduced.

Searching regularities have been intensively studied for many years and a lot of algorithms have been proposed. Many of the algorithms are not similar to any other, though they solve similar problems. It is also difficult to understand many of them. Moreover, many of the algorithms are difficult to extend, they solve only one specific problem. However, in this work, notion of finite automata is used to solve covers- and seeds-related problems. It is motivated by intuitive and elegant solution to various problems in stringology. The algorithms presented here directly follow from basic principles and properties of covers and seeds that are modelled using the formalism of finite automata. Therefore, the algorithms are easy to understand, similar to each other, straightforward and easy to implement or extend them to similar problems.

Organization of the Text

The central topic of the work is divided into two main chapters, which are closely related. First one discusses the problem of finding all covers of a string and introduces a solution, while in the second one the problem of finding all seeds of a string is discussed and the solution is brought.

In both main chapters, after description of all necessary properties of covers or seeds respectively, the algorithms for computing exact or approximate ones under Hamming distance are introduced. Next, the correctness and time and space complexities are investigated. Finally, some experiments demonstrating behavior of new algorithms under different input conditions or relations between various input parameters and input strings are performed and the results are discussed.

Advantages

Searching string regularities is a discipline among the borders of stringology. It may be important in many other fields of knowledge like data processing or molecular biology. This work is a valuable contribution to the problem of finding covers and seed of strings, especially approximate ones. Namely the algorithm for finding all k-approximate seeds, restricted or not, is the only algorithm solving this problem so far.

Despite the fact that the presented algorithms solve complex and difficult issues, they are very easy to understand and straightforward to implement due to the well-known formalism of finite automata as their bases and therefore usable for wide class of researchers. Since finite automata are studied for decades and there are numerous known extensions or modifications of them, it is also simply possible to extend the algorithms to solve different sorts of problems.

Disadvantages

Although all algorithms are clearly described and correctly investigated, performed experiments do not demonstrate them in the best way. Most of resulting graphs are created only above the small set of measured values, thus the relations of input variables are not always clear or cannot be even read. The set of input strings representing chromosomes was chosen appropriately to demonstrate one of the possible usages of the algorithms, however, it would be also beneficial to consider any other data set or field of usage, since the properties of the alphabet of chromosomes are very special. It would be also interesting to at least briefly compare finite automata approach to some other possible approaches.

Open Problems

All introduced algorithms deal only with Hamming distance, which is among others limited only to comparing strings of equal lengths. Therefore presented algorithms cannot work with distance functions allowing insertions or deletions of symbols. It would be certainly valuable extension of the present work to consider also some other distance metrics. Also as a distance while matching with the input string only the fixed number of errors is allowed, thus some other dynamic approach would be interesting for future research. Moreover, regularities can be searched not only in strings, but for example in tree structures, which could be also one way how to build on this thesis. Since presented algorithms work at least under cubic complexity in relation to the input string, it would be highly appropriate to find parallel versions of them.

Suggested Corrections of Minor Mistakes

The present PhD thesis contains several minor mistakes, such as various typos and misprints. Their corrections follow next.

1. page 1, paragraph 2, line 6 – correct "have" to "has"
2. page 1, paragraph 3, line 1 – correct "have" to "has"
3. page 2, line 15 – add "and" into "seed of w and a factor"
4. page 4, section 1.4, line 9 – change "one" to "ones"
5. page 8, last line – correct "inteterminate" probably to "indeterminate"
6. page 14, Note 2.4.22 – add "of" into "A left seed of a string"
7. page 15, line 2 – remove multiple occurrence of "automaton"
8. page 15, Definition 2.5.8 – correct to "sequence of transitions"
9. page 16, Definition 2.5.18 – correct to "sequence of transitions"
10. page 17, Definition 2.5.22, line 3 – to be consistent, the definition of the language should be
    L = { w: w \in A*, q \in \delta*(q_0,w), q \in F}
11. page 17, Definition 2.5.26, line 2 – correct to "they accept"
12. page 21, line 6 – add "of" into "focused on use of the formalism"
13. page 21, line 7 – correct "contribution of" to "contribution to"
14. page 24, section 3.1.3, line 1 – correct "automata ... is" to "automata … are" and reconsider word order of the sentence
15. page 25, Example 3.1.25, line 3 – remove multiple occurrence of "as"
16. page 26, line 11 – there is no source of nondeterminism, constructed automaton is merely incomplete
17. page 30, paragraph 4, line 2 – change "distance are" to "distances are"
18. page 30, last line – change "if a reversion" to "is a reversion"
19. page 36, line 6 – change "character-by character" to "character by character" or "character-by-character"
20. page 41, Lemma 4.1.8, proof, only if, line 4 – remove bracket
21. page 46, line 9 – change "the Lemma" to "the lemma" or "Lemma 4.2.9"
22. page 46, Theorem 4.2.10, proof, line 6 – add "are" into "there are at most … elements"
23. page 48, last line of the proof – change "the Lemma" to "the lemma" or "Lemma 4.2.16"
24. page 50, last line of the first proof – change "the Lemma" to "the lemma" or "Lemma 4.2.20"
25. page 57, Theorem 4.2.32, proof – change "the Algorithm" to "the algorithm" or "Algorithm 4.3"
26. page 58, Note 4.2.36, line 3 – change "the Lemma" to "the lemma" or "Lemma 4.2.35"
27. page 63, paragraph 3, line 7 – correct "The of dependence of …" to "The dependence of …"
28. page 70, bottom left graph – omit some numbers to make the description of the bottom axis more readable
29. page 75, top graph – make the description of the bottom axis more readable
30. page 79, paragraph 3, last sentence – omit commas before "of"
31. page 92, Proposition 5.2.27, proof, line 6 – change "the Algorithm" to "the algorithm" or "Algorithm 5.1"
32. page 93, last line of the first proof – change "the Lemma" to "the lemma" or "Lemma 5.2.29"
33. page 94, Theorem 5.2.36, proof, line 4 – reference (5.2) is either algorithm or chapter, not a number
34. page 95, section 5.3.1, paragraph 2, line 5 – change "The dependence on the number of states on k …" to "The dependence of the number of states on k …"

35. page 95, section 5.3.1, paragraph 2, line 6 – add dot before "Another"
36. page 95, section 5.3.1, paragraph 3, line 9 – change "depending … distance" to "depending on … distance"
37. page 100, line 7 – change "The dependence on the number of states on k …" to "The dependence of the number of states on k …"
38. page 100, line -3 – change "depending … distance" to "depending on … distance"
39. page 108, Chapter 5, last sentence – chapter 5 is not about covers, change "covers" to "seeds"

Conclusion

The topic of the PhD thesis is very important. In essence, the thesis has accomplished all its goals. The methods applied in the thesis are appropriate. This work represents a significant contribution to computer science and informatics, such as language and text processors or chromosome processing. It contains valuable algorithms, which above that solve so far unsolved problems. Therefore, this thesis satisfies all the PhD requirements, so

I recommend the PhD thesis by Ondřej Guth for its defence.

Prof. RNDr. Alexander Meduna, CSc
FIT BUT, Brno

Report on the Doctoral Thesis: Searching Regularities in Strings using Finite Automata

Author: Ing. Ondřej Guth

Searching regularities in strings, i.e. words over a given alphabet, constitutes an important research topic in computer science and its applications. Let us mention e.g. applications in molecular biology (searching regularities in chromosomes), applications in computer science in data compression, speech recognition, and others. Even though regularities in strings were intensively studied during the last decades, many questions remain open. The proposed thesis brings a new and unifying approach as well as new results by employing finite automata for designing algorithms that construct covers, approximate covers, seeds, and approximate seeds of a given string.

The thesis is divided into six chapters. The first two are devoted to the problem formulation, motivation, and definitions of all notions used further in the text. In Chapter 3, known results concerning covers and seeds are summarized; known algorithms that are furher used are also presented.

The main new results are contained in Chapter 4 and Chapter 5. Chapter 4 brings results concerning covers (exact and k-approximate with respect to Hamming distance), Chapter 5 is devoted to solving similar problems for seeds (exact and k-approximate with respect to Hamming distance). In Chapter 4, algorithms that constructs all covers of a given string are presented: at first for exact covers, then for k-approximate covers with respect to Hamming distance. For all algorithms their correctness is proven, and both time and space complexities are estimated. Finally, the new algorithms are tested: the number of covers that are computed with respect to the length of a tested string, and the time and space requirements of the algorithms, again with respect to the length of the tested string. Chapter 5 bring similar results for seeds.

The thesis ends with conclusions (Chapter 6) where also possibilities of further research are mentioned.

The thesis is well structured and written, even though it is not void of some mistakes and typos. The style is clear, the concepts and results are presented in an adequate way. Algorithms are illustrated on small examples which helps the reader. The author presents his ideas clearly. The fact that all notions are gathered in Chapter 2 would be very helpful if the notions could serve as links, in a printed form it is less useful. Because of that, I would appreciate more detailed references in the subsequent chapters. The bibliography is broad, thorough and current. The thesis are written in English which is adequate and does not make the reading of the thesis difficult.

I have the following remarks and comments:

- Definition of the extended transition function of an ε-NFA is not correct (Definition 2.5.21 together with Definition 2.5.22).
- When calculating the number of states of the cover-candidates automaton the upper bound is unnecessarily big. The upper bound for $\sum_{i=0}^{n} i^k$ is $\mathcal{O}(n^{k+1})$ (better than $\mathcal{O}(n^{2k})$ given in the thesis). Better estimates can be obtained also in Theorem 4.2.25 and Note 4.2.26.
- When calculating time and space complexity for seeds, formula (5.2) should be
$$|\mathbf{w}| \cdot k \cdot \binom{|w|}{k} \cdot \frac{(|A|-1)^{k+1}-1}{|A|-2}.$$
This will also change estimates in Theorems 5.2.34 and 5.2.35.
- Proposition 4.2.22 should be formulated only for the Hamming distance (and not for a general distance function), otherwise it needs a thorough new proof.
- The author sometimes does not distinguish between a state and its depth (e.g. page 43, 52, 53, 83).
- Notation of states in Algorithms 5.2 and 5.4 is not very well-chosen; it mixes a state and its level. (See line 11 in Algorithm 5.2, lines 3 and 4 in Algorithm 5.4).

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, 18. 7. 2014

M. Demlova

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

Examiner's Report: Ph.D. Dissertation

"Searching Regularities in Strings using Finite Automata"

Ondřej Guth

Czech Technical University in Prague

W. F. Smyth, McMaster University

June 10, 2014

W. F. Smyth

1 Relevance

Over the last 25 years, stringologists have come to realize that the notion of periodicity in strings needs to be relaxed and extended, so that other "regularities" can be quantified and exploited in algorithms to discover new structures in strings. See for example [2, 1].

2 Formal Structure & Organization

The dissertation is exceptionally well organized, with three introductory chapters, the first (Introduction) giving an overview of the results, the second (Preliminaries) providing all the fundamental definitions, the third (Background & State-of-the-Art) presenting a comprehensive exposition of current research in string regularities.

3 Completion of Objectives

In the abstract, the author states that the "main contributions of this thesis are algorithms for computing all approximate covers or seeds of a string based on formalism of finite automata". This goal is certainly achieved.

4 Assessment of Methodology

The methodology is based on the ideas of Voráček [3], who describes algorithms that use finite automata to compute all the covers of a generalized (indeterminate) string. In Chapter 4 the candidate extends these ideas to compute, for a given string $\boldsymbol{w}$, all the smallest distance k-approximate covers of $\boldsymbol{w}$ (under Hamming distance k), as well as all the *restricted* smallest distance k-approximate covers. In Chapter 5 these algorithms are extended still further to handle the same two computations for k-approximate seeds. To my knowledge these results are new: no previous algorithms had been found, whether using finite automata or not, to solve these difficult stringological problems. This outcome is all the more impressive, since generally the computation of seeds is much more difficult than the computation of covers: using finite automata, the author has found a way to achieve this generalization while using essentially the same approach to both categories of problem.

5 Evaluation of Results & Contributions

These results are a significant contribution to the computation of regularities in strings: the introduction of the k-approximate feature is particularly useful, since it extends greatly the notion of "quasiperiodicity", and does so in a way that makes it possible to relate this concept to a much wider range of strings. There may well be practical applications to bioinformatics, where sections of genome can be approximately copied many times to adjacent positions, thus yielding a sequence of DNA that is "approximately covered".

6 Remarks & Questions for the Defence

1. One obvious question that arises relates to the possibility of extending these results to edit distance. Such an extension would greatly enhance the applicability to bioinformatics.

2. Does the candidate envisage extension of these ideas to computation of an approximate cover array? Or to any of the cover array extensions considered in [1]?

3. The worst-case time complexities of the new algorithms are $\mathcal{O}(n^2)$ or higher. No claims are made that the algorithms are asymptotically optimal. Does the candidate have any ideas for possible reduction in execution time?

4. The use of English in the dissertation is neither standard nor colloquial, but it is always understandable — the applicant is to be congratulated for writing so clearly in a foreign language.

7 Overall Evaluation

The author of the dissertation has demonstrated the ability to conduct research and achieve scientific results. In accordance with para. 47, letter (4) of Law 111/1998 (the Higher Education Act), I recommend this thesis for presentation and defence for the award of the Ph.D. degree.

References

[1] Tomáš Flouri, C. S. Iliopoulos, Tomasz Kociumaka, Solon P. Pissis, Simon J. Puglisi, W. F. Smyth & Wojciech Tyczyński, **Enhanced string covering**, *Theoret. Comput. Sci.* 506–30 (2013) 102–114.

[2] W. F. Smyth, **Computing regularities in strings: a survey**, *European J. Combinatorics* 34–1 (2013) 3–14.

[3] M. Voráček, *Algorithms on Generalized Strings*, dissertation thesis proposal, Czech University in Prague (2005).