PhD Thesis Evaluation

Title: Automata with Counting in Regular Expression Matching

Author: Ing. Lenka Turoňová

1 Thesis overview

Regular expressions (regexes) are a widely used formalism for pattern matching and text processing supported by many tools and programming languages. Although regexes are very useful, they may cause computational problems resulting from their high worst-case computational complexity. In particular, an attacker could use the knowledge of the implementation to significantly slowdown the regex matching, and consequently the provided service. Such attacks are known as Regular expression Denial of Service (ReDoS) attacks. In her thesis, Ing. Turoňová focuses on this kind of attacks.

The thesis is divided into six sections. The first section introduces the problem of ReDoS attacks, and briefly overviews their history and existing approaches to their detection. It emphasises that the main focus of the thesis are regexes with bounded repetition. The second section is devoted to the introduction of the necessary notions and concepts. The main contributions are presented in Sections 3–5.

Section 3 studies regexes with counting in the context of ReDoS attacks for both backtracking and automata-based matchers. The attack for automata-based matchers, which is the main focus of this section, comes basically from the high cost of the determinization of an NFA constructed from a given regex. To tackle this problem, the thesis proposes *counting automata* (CAs) and an algorithm to construct a deterministic CA that is exponentially more succinct compared with the corresponding DFA. It further discusses a special class of CAs – monadic CAs – corresponding to regexes with bounded repetition restricted to character classes. Although monadic CAs (or regexes with bounded repetition restricted to character classes) form a subclass of general CAs, the study of this subclass is justified by practical experiments, where monadic regexes form over 95 % of the considered dataset.

Section 4 further improves the study by introducing a notion of a *counting-set automaton* (CsA), which is a finite automaton with a simple, but very clever data structure. Rather than determinized, a CA is converted to a deterministic CsA in time that does not depend on the repetition bounds of the regex. The simulation time of the resulting CsA is linear in the size of the input text. A disadvantage of this approach may seem to be that it covers only a subclass of regular expressions. However, this subclass is experimentally shown to cover the majority of regexes used in practice. An extensive experimental study then supports the advantage of the use of CsAs.

Section 5 systematically studies regexes with counting in the context of ReDoS attacks for automata-based matchers. This part proposes an evil-text generator and experimentally shows that the generator is more "evil" than the existing generators.

Finally, Section 6 summarizes the contributions and suggests future research directions.

2 Summary of publications

The thesis is based on three publications. Although the publications are coauthored by six authors, Ing. Turoňová has significantly contributed to all of them. In particular, Section 3 is based on "Holík, Lengál, Saarikivi, Turoňová, Veanes, Vojnar. Succinct Determinisation of Counting Automata via Sphere Construction, APLAS 2019", where Ing. Turoňová contributed 50 % and presented the paper, Section 4 is based on "Turoňová, Holík, Lengál, Saarikivi, Veanes, Vojnar. Regex Matching with

Counting-Set Automata. Proc. of the ACM on Programming Languages 4(11), 1–30, 2020", where Ing. Turoňová contributed 60 % and presented the paper at OOPSLA 2020, and Section 5 is based on "Turoňová, Holík, Homoliak, Lengál, Veanes, Vojnar. Counting in Regexes Considered Harmful: Exposing ReDoS Vulnerability of Nonbacktracking Matchers. USENIX Security 2022", where Ing. Turoňová contributed 60 %. Other publications of the author are:

1. Horký, Sič, Turoňová, Holík, Automata with Bounded Repetition in RE2. EUROCAST 2022, contribution 30 %
2. Janků, Turoňová, Solving String Constraints with Approximate Parikh Image. EUROCAST 2019, contribution 50 %
3. Holík, Turoňová, Towards Smaller Invariants for Proving Coverability. EUROCAST 2017, contribution 50 %

3 Evaluation

The thesis consists of significant contributions to the problem of ReDoS attacks from both theoretical and practical point of view. Theoretical contributions consist in the introduction of two types of counting(-set) automata, a clever data stucture, and determinization algorithms, while the practical contribution consists in an extensive comparison of the newly invented techniques with existing techniques and tools, all based on a large dataset of real-world regexes.

4 Questions and comments

1. Derivatives are closely related to the coalgebraic view on finite automata and there are coalgebraic minimization and determinization algorithms (e.g., Adámek et al., A Coalgebraic Perspective on Minimization and Determinization, FoSSaCS 2012). Could it be useful in your study?
2. You define the size of a regex, denoted by $|R|$ and later by $\#(R)$, as the number of nodes in the abstract syntax tree. Ellul et al., Regular Expressions: New Results and Open Problems, J. Automata, Languages and Combinatorics 9, 2004, 233–256, have discussed the size of regexes and claim that the best choice is to use the number of alphabetic symbols with repetitions. Considering this definition of the size of regexes, how do these definitions compare and would this definition have any significant impact on your results?
3. You use Antimirov's construction to build an NFA from a regex. Have you considered other constructions such as, e.g., that of Hromkovič et al., Translating Regular Expressions into Small ε-Free Nondeterministic Finite Automata, J. Comput. System Sci. 62, 565–588? Why did you choose just Antimirov's construction?
4. Example 3.4.1. could be more detailed, also the proofs in Section 3 are only sketches. Why?
5. The experimental results in Section 3.5 are very interesting and impressive. I am just wondering, was "DFA ever better than Counting" in your experiments?
6. Section 4: Do you think you could extend the counting-set (or similar) idea to general regexes?
7. On page 66 and in Example 4.4.4, you give an example of a regex you cannot handle with CsA. Let's call the regexes that have a uniform CA uniform. What can you say about non-uniform regexes? In particular, are all non-uniform regexes equivalent to a uniform one? Are there regexes that are not equivalent to any uniform regex?
8. In Figure 4.7, the last expression seems to be pretty bad for CsAs. Could you comment on this?
9. On page 95, you compute the weight of states as the maximum of the weight of the state and the weight of its successor minus 0.5. Could you explain in more detail why "minus 0.5"? Would a different constant make a difference?
10. What does "node" in Table 5.1 stand for?

11. On page 104, you say "from 22.425 original regexes we removed 16.094 regexes not supported by our tool..." it looks like your tool does not support most of the regexes? Could you comment more on what are the eliminated regexes?
12. Table 5.7, the first regex seems not to contain counting. What makes it so hard to cause such a slowdown?

A list of a few typos:

- Page 9, ref. [58] is dated 1951 and not 1956.
- The definitions of derivatives are redundant, namely, the first two Brzozowski derivatives are covered by the third (there is even an inconsistency between the second and third), similarly for Antimirov derivatives.
- Ref. [13] at the end of page 18 should be [4].
- Last line of the proof of Theorem 2, $|\Sigma|$ should probably be removed.
- The counting-set data structure is very clever. However, in the definition, what do you mean by $\ell := 0$? Isn't ℓ a queue?
- Several typos on pages 51 and 53, mainly a lower-case "t" after the period.
- Page 54, definition of $\partial_\alpha(XZ)$: should the first S on the right-hand side be rather X?
- Some references to the text are wrong, such as on page 56, "...the predicate shown above" is actually below. Similar problem appears several times.
- There are no Lemma 4.13 and Theorem 4.14.
- Section 80?
- Page 95, should $\{(6, c = 2)\}$ be $\{(6, c = 1)\}$?
- Page 96, Figure 5.1? It is confusing to refer to Section 5.3 within Section 5.3.
- Section 6.1, "formata-based"?
- Some references do not contain all the details, e.g., [13] does not contain the name of the journal.

5 Conclusion

The candidate, Ing. Lenka Turoňová, demonstated creative abilities, and her PhD thesis presents original and important research results. I recommend that the candidate is awarded the PhD degree.

June 6, 2022 doc. RNDr. Tomáš Masopust, PhD., DSc.