UDC 004.02
IRSTI 20.53.15

EXTENDED REGULAR EXPRESSIONS IN FINITE AUTOMATA REVISITED

Mirzakhmet Syzdykov

al-Farabi Kazakh National University, Almaty, Kazakhstan
mspmail598@gmail.com
ORCID ID: https://orcid.org/0000-0002-8086-775X

Abstract. While the past work was focused on limited set of regular expressions (RE) within extended operators like intersection, complement and subtraction (ERE), in this article we extend the definition of RE for zero-width operators like Kleene closure. For this purpose the tagged states are implemented within the space of states of the non-deterministic finite automata (NFA) as well as for modified subset construction of deterministic finite automata (DFA). The prior research is good for extended operators, however, empty words play vital role even in extended regular expression matching as well as for typical regular expressions. Thus, the tagged states and transitions are introduced as well as the local search, which was first developed for approximate back-reference matching and now is suitable for extended operators. Thus, the linear complexity isn't avoided and is obtained for the general case of the grammars of regular expressions. It's also shown that the limited set of grammar rules for regular expressions are the good idea to obtain preliminary results for further generalization using algorithmic paradigms.

Keywords: extended operators, regular expression, algorithm.

Introduction

The prior work for ERE has the results as the membership problem for intersection is closed under LOGCFL [1] and are quadratic in general to the number of states in produced NFA.

Berry and Sethi propose derivative within the extended operators [2], however, the state explosion in this case cannot be avoided.

Neven et al. [3] study the extended operators and show comparable results to the previous research.

Laurikari, Ville, [4] introduces the tagged transitions in NFA – this is the basic idea of our present approach which will be described later in this article.

Rabin-Scott [5] construction within the extended operators [6] is also presented by the algorithm with tagged transitions.

We will use here also the local search algorithm as an alternative to the tagged transitions and show that it doesn't give the quadratic explosion of complexity or space-time tradeoff [7].

The latest research was based on limited set regular expressions [8] while the Kleene star operator was missing. In this article we show the general case when zero-width operators are included within the local search in NFA or within the modified subset construction on tagged NFA.

The best known result for the complexity of extended regular expression matching is presented in [11]. We out-perform in this article this bound giving the complexity of O(mn) which is linear to both factors like the size of the expression as well as the size of the input parameter.

The problem of finding the equivalence for the short regular expression along the intersection operator is presented in [12]. Thus, the equivalence of languages can be also solved efficiently by our algorithm.

Definition of Limited ERE. Our prior research was mainly focused on limited set of non-zero width operators in ERE which are as follows:

L(R1 | R2) = L(R) : w in L(R1) or w in L(R2),
L(R1 & R2) = L(R): w in L(R1) and w in L(R2),

L(R+) = L(R): w in L(R)+,
L(R1 – R2) = L(R): w in L(R1) and w not in L(R2),
L(~R) = L(R): w not in L(R).

The above definition was used in the prior works of the author [6, 7, 8] while the Kleene star operator and zero-width elements were missing which is thus the limited set of regular expressions. Of course, we can insist on the fact that they are sufficient for matching process and are good alternative as for ideas proposed by Dominik D. Freydenberger [9] of the limited sets.

However, to feel the full power of regular expressions the zero-width Kleene closure as well as the empty word is to be introduced as they are widely used in practice to represent the common sets within the pre-defined task of the operator to code the language or pattern for matching against input string.

The limited REs presented in this article, however, aren't limited to the extended operators like intersection, subtraction and complement: at least the empty words are missing. This will be discussed further.

Definition of Zero-width ERE. The zero-width ERE mainly consist of Kleene closure operator and include empty words which are to be handled correctly without loss of generality. We've already mentioned that our prior research focused on limited set of ERE.

Thus, zero-width ERE are as follows:

L(R*) = L(R): w in L(R)*,
L(R?) = L(R): w in L(R) or w is an empty word,
L(empty word) = empty word.

Thus, we extend our common set of regular expression grammars to the zero-width operator like Kleene closure. As per the prior work [6, 7, 8] it was the limited set, however, later in this article we will show how to cope with the solutions and avoid complexity and code expansion.

The star operator gives us the possibility to encode the empty words as an alternative to the either finite or infinite sets of words in language L(R).

Zero-width ERE can be found in almost all regular expression flavors to the present time and cannot be posed as the limited set as we used it for an initial evaluation of the development of the algorithms for the general case within the operators above.

Tagged NFA for Modified Subset Construction

We tag the starting state in our construction and there onwards we save it in the state, thus making the construction possible. Another alternative to this approach would be usage of local search for NFA matching, however, our proposal is for subset construction by Dana and Scott.

In this case the quadratic explosion is avoided due to the fact that the produced DFA can be stored in table-like form.

The tag is to be placed on the incoming state from NFA construction for ERE which is defined in [6]. Thus, the tag traces the operability of the finishing state and avoids errors within the nature of concatenation operator which is essential as well as for RE and as for ERE.

Tagged NFA was introduced in [4] and was used for efficient submatch addressing, in our work we use tagged NFA with free marks to track the reachability for extended operators like intersection, subtraction and re-written complement.

The modified subset construction which was first introduced for intersection operator in [10] was of limited nature for regular expressions, however, along with the latest research presented in this article, it's clear that tagged transitions as well as tagged states in NFA solve the general case problem for zero-width operators like star (*) or Kleene closure operator.

We have to note that complexity remains the same as per limited set of operators [8] in NFA construction which is discussed further in the next section.

In any case the fact that the pre-computation is linear and the produced result is also linear

gives us the better understanding of the future of ERE matching realization from the algorithmic point of view.

Local Search and Tagged NFA alternative

The scalable window approach opposite to the naive method uses flowing window event when the The better alternative for NFA would be approximate back-reference matching for ERE [7]. In this case the linear bound for complexity also holds true as we save the local search addendums in each phase of iteration. The O(mn)-memory complexity is also true for tagged NFA for proper matching.

As we can see the memory consumption in this case cannot be avoided while the running time complexity goes for a better side in O(mn), where m is the size of the pattern and n is the size of the sought string.

For the local search it's impossible to construct the DFA via subset construction, however, it's operable for extended regular expressions on NFA and is very effective against the quadratic complexity explosion.

The local search is also linear as well as the alternative tagged NFA construction and matching as we have to go further for the whole input, however, this fact gives us the notion for on-line algorithms which operate on the timely limited set of data. Thus, the whole string for matching is to be defined for local search rather than on-line perspective of the past work [6, 7, 8].

The local search [7] used for back-reference matching can be without loss of generality also used in ERE matching – however, this is not so simple and challenging task for practical evaluation. We can note here that limited set of regular expressions is a good solution to avoid the bad practices for algorithmic design and composition.

Conclusion

We have introduced the main goals of our experimentation with extended operators and original automata to obtain better results for ERE matching. Now, it's time for realization which is algorithmic in main case within the modified subset construction for zero-width Kleene closure or tagged NFA matching, or its alternative like local search.

Thus, the tagged NFA and local search alternatives are proposed for both NFA-matching or modified subset construction without loss of generality for star operator and empty words, which together form the practically full set of regular expressions with the extended operators like intersection, complement and subtraction.

We also use the paradigm of algorithmic simplicity which is related to the realization process, thus obtaining the efficient and linear algorithm which is well-studied in short for a better evaluation. The Java project from repository can be obtained from the author by demand.

Thus, we state the new linear complexity of two factors in O(mn), where m is the size of the regular expression pattern and m is the size input string to be matched. The memory consumption due to the tagged NFA and produced DFA or local search in NFA remains same.

We also obtain a good result of not-rewriting the union operator like in prior author's work using the methods developed in this work.

Acknowledgements

The author expresses gratitude to the Professor Steven Kearns for his valuable comments to make this work possible. The major test cases were proposed by him.

The author also expresses gratitude to Professor Dominik D. Freydenberger for his valuable ideas about limited regular expressions.

Funding

This work was partially supported by an educational grant of the Ministry of Education and Sciences of Republic Kazakhstan during author's work at the Institute of Problems in Informatics and Control (IPIC) from 2006 to 2009.

References

[1] Petersen, Holger. The membership problem for regular expressions with intersection is complete in LOGCFL. Annual Symposium on Theoretical Aspects of Computer Science. Springer, Berlin, Heidelberg, 2002.

[2] Berry, Gerard, and Ravi Sethi. From regular expressions to deterministic automata. Theoretical computer science. 1986. 48. 117-126.

[3] Martens, Wim, Frank Neven, and Thomas Schwentick. Complexity of decision problems for simple regular expressions. International Symposium on Mathematical Foundations of Computer Science. Springer, Berlin, Heidelberg, 2004.

[4] Laurikari, Ville. NFAs with tagged transitions, their conversion to deterministic automata and application to regular expressions. Proceedings Seventh International Symposium on String Processing and Information Retrieval. SPIRE 2000. IEEE, 2000.

[5] Rabin, Michael O., and Dana Scott. Finite automata and their decision problems. IBM journal of research and development 1959. 3(2). 114-125.

[6] Syzdykov, Mirzakhmet. Deterministic automata for extended regular expressions. Open Computer Science. 2017. 7(1). 24-28.

[7] Syzdykov, Mirzakhmet. Local search in non-deterministic finite automata with extensions. J. Advanced technologies and computer science. 2021. 3. 24-28.

[8] Syzdykov, Mirzakhmet. Membership Problem in Non-deterministic Finite Automata for Extended Regular Expressions in Linear Polynomial Time. J. Advanced technologies and computer science. 2021. 4. 14-17.

[9] Freydenberger, Dominik D., and Markus L. Schmid. Deterministic regular expressions with back-references. Journal of Computer and System Sciences. 2019. 105. 1-39.

[10] Syzdykov, Mirzakhmet. Algorithm to Generate DFA for AND-operator in Regular Expression. International Journal of Computer Applications. 2015. 124(8).

[11] Kupferman, Orna, and Sharon Zuhovitzky. An improved algorithm for the membership problem for extended regular expressions. International Symposium on Mathematical Foundations of Computer Science. Springer, Berlin, Heidelberg, 2002.

[12] Gelade, Wouter, and Frank Neven. Succinctness of the complement and intersection of regular expressions. ACM Transactions on Computational Logic (TOCL). 2012. 13(1). 1-19.

Сведения об авторе:

Анг.: Syzdykov Mirzakhmet - al-Farabi Kazakh National University, Almaty, Kazakhstan

Каз.: Сыздықов Мырзахмет- әл-Фараби атындағы Қазақ ұлттық университеті, Алматы, Қазақстан.

Рус.: Сыздыков Мырзахмет- Казахский национальный университет имени аль-Фараби, Алматы, Казахстан.