2023.6.16

NTT Corporation

NTT and Waseda University develop novel technology for automatically repairing regular expression error of the extraction function

NIPPON TELEGRAPH AND TELEPHONE CORPORATION (NTT, Head Office: Chiyoda-ku, Tokyo; President & CEO: Akira Shimada) and Waseda University have developed practical novel technology for automatically repairing bugs of regular expressions (regexes) for extraction that slow down web applications or cause information leakage, in the worst case, by exploiting the behavior of regular expression engines. Regular expressions (*1) are patterns used to match character combinations in strings and are used for a wide variety of web services to sanitize user inputs, extract data, etc. Regular expressions can describe complex string patterns concisely. On the other hand, regular expressions are challenging to understand in detail, and cases have been reported in which incorrect regular expressions remain without repair.

This technology allows developers who do not have expertise in regular expressions to repair bugs of regular expressions and helps to provide secure services.

Details of this technology will be presented at PLDI2023 (*2), one of the most prestigious international conferences in the programming languages field, held from June 17 to 22, 2023.

1. Background

Regular expressions are built into and widely used in most programming languages and are extensively utilized in various software/services, such as for extracting user IDs from website URLs (Fig.1). However, accurately understanding the behavior of programs using regular expressions can be challenging for humans, and incorrect regular expressions still need to be corrected in publicly available open-source programs (*3). These wrong regular expressions can cause system malfunctions, leading to information leakage and service interruption. Moreover, a growing number of cyber-attacks deliberately exploit these vulnerabilities, posing a risk to stable service delivery.

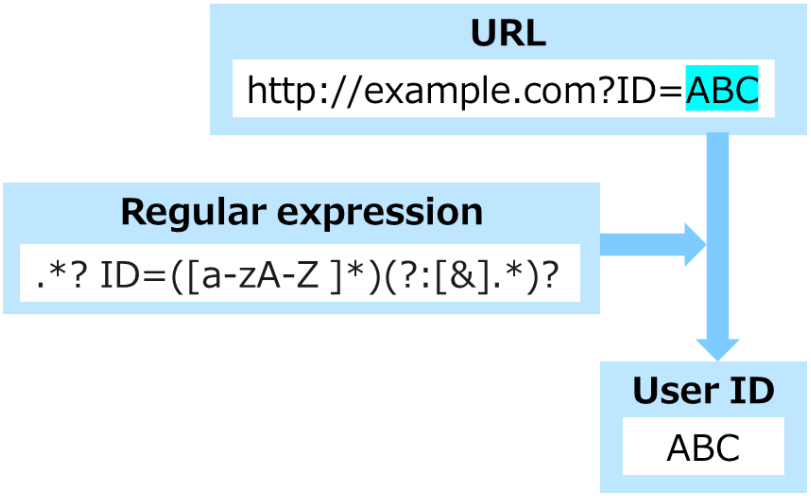

Figure 1. Examples of extraction using regular expression

(Example of extracting a string consisting of upper or lower case alphabetic characters after "ID=" and before the trailing or "&" character)

2. Achievement

The main applications of regular expressions include pattern matching in strings and extracting matched substrings from a text on the basis of intended patterns. However, extracting substrings is required to handle a greater variety of variations than pattern matching, and it requires a high level of expertise to accurately describe unambiguous regular expressions, and traditionally, mechanical repairs have been challenging to make.

In this project, we achieved the world's first technology for automatically repairing bugs of regular expressions for extraction by giving a first formal definition of actual regular expression engines, a repair problem, and an algorithm for solving the repair problem (Table 1). This technology allows developers who do not have expertise in regular expression to easily repair incorrect regular expressions, reducing risks such as information leakage and service interruptions.

NTT developed the definition of actual regular expression engines and the repair algorithm, and Professor Tachio Terauchi (Faculty of Science and Engineering, Waseda University) verified the theoretical correctness of the method developed by NTT.

Table 1. Positioning of this technology

		Automatic repair technology	
	Usage	Existing technology	Our technology
Regex	pattern matching	✓	✓
	Extraction	–	✓

3. Technology

The key points of this technology are as follows:

① The formal definition of programs that perform the matching of regular expressions (regular expression engines)

② The method for generating a condition that guarantees the correctness of the repair

③ The alorithm for repairing regular expressions for extraction from positive examples (strings to be accepted with information of substrings to be extracted) and negative examples (strings to be rejected). The algorithm uses the functionality to abstract a regular expression for reducing the running time and the functionality to make the abstracted regular expression concrete alternately (Fig. 2).

This technology uses the formal definition of regex engines that follows the ECMAScript 2023 language specification (*4), widely used in web applications.

This method verifies whether a program satisfies a formal specification and proves the correctness of the program in accordance with the specification. Such a method is called "formal verification(*5)". Although the correctness shown by the conventional verification in test cases is limited, the formal verification is comprehensive. It will help the users to develop secure software.

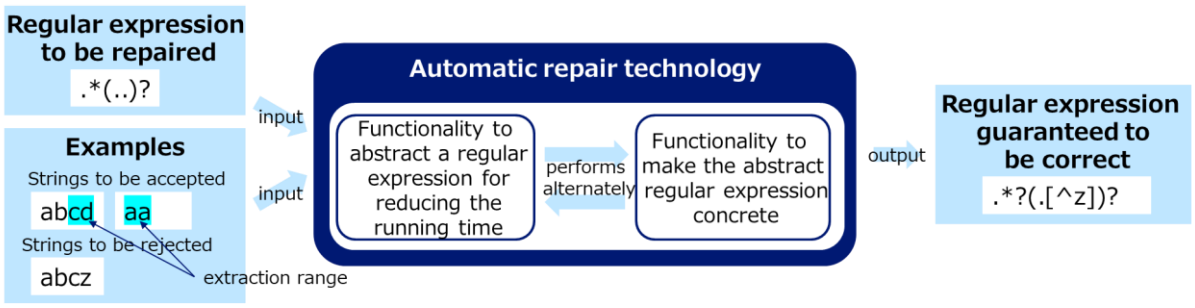

Figure 2. Examples of repairing regular expressions for extraction

4. Future plans

This technology allows developers who do not have expertise in regular expressions to repair bugs of regular expressions and helps to provide secure services.

Although the use of AI for generating programs is becoming common, a new issue has arisen regarding how to deal with errors in programs generated by non-experts using AI. Therefore, this technology is expected to improve program security without compromising the benefits of AI-driven automation.

About the announcement

This result will be presented at PLDI2023 (Programming Language Design and Implementation 2023),

one of the most prestigious international conferences in the field of programming languages and programming systems research, to be held from June 17 to 22, 2023, with the following titles and authors.

Title: Repairing Regular Expressions for Extraction

Authors: Nariyoshi Chida (NTT Social Informatics Laboratories), Tachio Terauchi (Waseda University)

URL: https://dl.acm.org/doi/10.1145/3591287

Reference

※1 Regular expressions

Method used to match character combinations in strings.

※2 PLDI

Programming Language Design and Implementation is a conference of ACM SIGPLAN and is the most prestigious international conference held in the field of programming languages and programming systems research, covering the areas of design, implementation, theory, applications, and performance.

URL: https://pldi23.sigplan.org/

※3

According to a survey of string-related software bugs, the open-source project under investigation had 204 bugs, 75 of which were said to be regular expression errors.

※4 ECMAScript 2023

ECMAScript is the official standard for the JavaScript programming language, widely used in web applications. This technology has been validated against ECMAScript, updated in 2023.

※5 Formal Verification

The method that verifies the correctness of a program by a formal specification and helps the users to develop reliable software.

<Inquiries regarding this press release>
Nippon Telegraph and Telephone Corporation
Service Innovation Laboratory Group
Public Relations
Email: nttrd-pr@ml.ntt.com