Scalable Large-Variance Clone Detection

Tasuku Nakagawa[1,a)] Yoshiki Higo[1,b)] Shinji Kusumoto[1,c)]

Abstract: A code clone (in short, clone) is a code fragment that is identical or similar to other code fragments in source code. Clones generated by a large number of changes to copy-and-pasted code fragments are called large-variance clones. It is difficult for general clone detection techniques to detect such clones and thus specialized techniques are necessary. In addition, with the rapid growth of software development, scalable clone detectors that can detect clones in large codebases are required. However, there are no existing techniques for quickly detecting large-variance clones in large codebases. In this paper, we propose a scalable clone detection technique that can detect large-variance clones from large codebases and describe its implementation, called NIL. NIL is a token-based clone detector that efficiently identifies clone candidates using an N-gram representation of token sequences and an inverted index. Then, NIL verifies the clone candidates by measuring their similarity based on the longest common subsequence between their token sequences. We evaluate NIL in terms of large-variance clone detection accuracy, general Type-1, Type-2, and Type-3 clone detection accuracy, and scalability. Our experimental results show that NIL has higher accuracy in terms of large-variance clone detection, equivalent accuracy in terms of general clone detection, and the shortest execution time for inputs of various sizes (1–250 MLOC) compared to existing state-of-the-art tools.

1. Introduction

A code clone (in short, clone) is a code fragment that is identical or similar to other code fragments in source code. Clones are generated by copying, pasting, and modifying code fragments for reuse. Clones are a major problem in software maintenance because they lead to bug propagation. Therefore, clone detection techniques, which automatically detect clones in the target codebase, are essential. Many clone detection techniques have been proposed [1].

It is important for clone detection techniques to detect clones that have been heavily edited. A clone generated by inserting or deleting a large number of statements in various places in a copy-and-pasted code fragment is called a large-variance clone. Such clones are common in software development and should thus be detected along with general clones. Wu et al. pointed out that it is difficult for existing clone detectors to detect large-variance clones; they proposed a technique for detecting such clones and presented its implementation, called LVMapper [2].

It is also important for clone detection techniques to be scalable. Highly scalable clone detectors are required for analyzing large-scale projects or source files in an inter-project repository. Many scalable clone detectors have been proposed [3, 4]. For example, to achieve scalable clone detection, SourcererCC [3] uses heuristics to reduce the number of code block comparisons needed to detect clones.

However, clone detectors that can detect clones with a large number of edits fail for large inputs [5] or require a long time to detect clones [2]. Scalable clone detectors target only identical or strongly similar clones (near-miss clones). They are incapable of detecting large-variance clones, in which many statements have been inserted or deleted. Therefore, the scalable detection of large-variance clones is challenging.

In this paper, we propose a scalable technique for detecting large-variance clones and describe its implementation, called NIL[*1], which uses an N-gram representation, an inverted index, and the longest common subsequence (LCS). NIL is a token-based clone detector. One of the features of large-variance clones is that the order of many tokens is preserved (i.e., the common subsequence between token sequences of large-variance clones is long). Hence, to detect large-variance clones, NIL measures the similarity between the token sequences of two code fragments based on the LCS. In addition, large-variance clones share many consecutive tokens. Hence, for scalable clone detection, NIL uses an N-gram representation of token sequences and an inverted index to reduce the number of code block comparisons needed to detect clones. First, NIL transforms code blocks extracted from source files into token sequences and creates an inverted index from the N-gram representation of the token sequences. Next, it identifies the clone candidates for each code block using the code block and the inverted index. Finally, it verifies the clone candidates by measuring the similarity between the code block and the clone candidates.

[1] Osaka University
Suita, Osaka 565–0871, Japan
[a)] t-nakagw@ist.osaka-u.ac.jp
[b)] higo@ist.osaka-u.ac.jp
[c)] kusumoto@ist.osaka-u.ac.jp

[*1] A clone detector using N-gram, Inverted index, and LCS. NIL is available at `https://github.com/kusumotolab/nil`.

```
protected String getPrompt(InputRequest request) {
  String prompt = request.getPrompt();
  if (request instanceof MultipleChoiceInputRequest) {
    StringBuffer sb = new StringBuffer(prompt);
    sb.append("(");
    Enumeration e = ((MultipleChoiceInputRequest) request)
       .getChoices().elements();
    boolean first = true;
    while (e.hasMoreElements()) {
      if (!first) {
        sb.append(",");
      }
      sb.append(e.nextElement());
      first = false;
    }
    sb.append(")");
    prompt = sb.toString();
  }
  return prompt;
}
```

(a) Clone A

```
protected String getPrompt(InputRequest request) {
  String prompt = request.getPrompt();
  String def = request.getDefaultValue();
  if (request instanceOf MultipleInputChoiceRequest) {
    StringBuilder sb = new StringBuilder(prompt).append("(");
    boolean first = true;
    for (String next : ((MultipleInputChoiceRequest) request)
      .getChoices()) {
      if (!first) {
        sb.append(",");
      }
      if (next.equals(def)) {
        sb.append('|');
      }
      sb.append(next);
      if (next.equals(def)) {
        ab.append('|');
      }
      first = false;
    }
    sb.append(")");
    return sb.toString();
  }
  else if (def != null) {
    return prompt + "[" + def + "]";
  }
  else {
    return prompt;
  }
}
```

(b) Clone B

Fig. 1: Example of large-variance clones

We evaluate NIL's (1) large-variance clone detection accuracy, (2) general Type-1, Type-2, and Type-3 clone detection accuracy, and (3) scalability. We compared NIL with existing state-of-the-art tools, namely LVMapper [2], CCAligner [5], SourcererCC [3], and NiCad [6]. The experimental results show that NIL has a high precision of 87% in large-variance clone detection. It also has a high recall of 100%, as determined in our evaluation of large-variance clone detection using a mutation technique. In general clone detection, the accuracy of NIL is equivalent to that of the existing tools. In addition, we confirmed that NIL has high scalability; it can detect clones faster than the existing tools for large inputs (codebases with 250 MLOC).

2. Preliminaries

2.1 Definition

A code fragment is a consecutive segment of source code. It can be represented by the tuple $(file\_name, start\_line, end\_line)$. A code block is a code fragment within braces. This study treats a function, which is a code block, as a clone detection unit, as done in previous studies [2,3,5]. Clones are code fragments identical or similar to other code fragments in source code. A pair of similar code fragments is called a clone pair. Clones are classified based on the degree of the similarity between them as follows.

Type-1 is an exact copy without modifications (except for white space and comments).

Type-2 is a syntactically identical copy; only variable, types, or function identifiers are different.

Type-3 is a copy with further modifications; statements have been changed, added, or removed.

The minimum length of clones is the minimum number of lines that a code fragment must be to be treated as a clone. It is often set to six lines or 50 tokens [1].

2.2 Large-variance clone

A large-variance clone is a clone generated by inserting or deleting a large number of statements in various places in a copy-and-pasted code fragment. Figure 4 shows an example of large-variance clones. In this example, statements have been inserted into and deleted from various places in Clone A to create Clone B. Wu et al. pointed out that existing clone detectors are incapable of large-variance clone detection because most target to the detection of near-miss clones [2]. They defined large-variance clone as follows. Consider two code blocks c_1 and c_2 with LOC values of L_1 and L_2, respectively, where $L_1 \leq L_2$. Let $\lambda = L_i/L_j$ (i.e., λ is the ratio of the code lengths of two code blocks). If c_1 and c_2 are Type-3 clones and the corresponding $\lambda \leq 0.7$, then these clones are large-variance clones. They proposed LVMapper, a large-variance clone detector.

LVMapper detects clones using code windows ((a code fragment composed of k consecutive lines in a code block)). Its clone detection has three phases, namely the locating, filtering and verifying phases. In the locating phase, LVMapper identifies pairs of code blocks that share at least one code window as clone candidates. Then, in the filtering phase, it calculates the proportions of common code windows for each clone candidate and removes clone candidates whose proportions are lower than filtering threshold θ. Finally, in the verifying phase, it verifies each clone candidate based on similarity measured using a common subsequence of lines between each clone candidate's code block pair.

3. Approach

Figure 2 shows an overview of the proposed technique. The input is a set of source code files, and the output is the clone pairs in the source code. In the proposed technique, large-variance clones are detected based on the similarity between token sequences based on the LCS, taking advantage

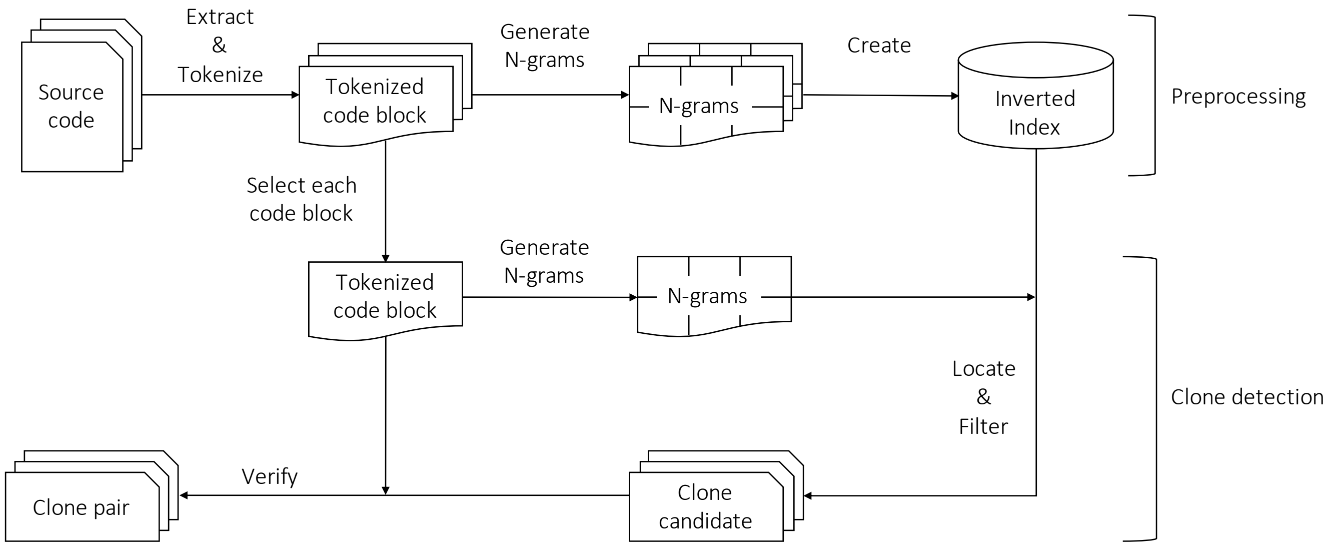

Fig. 2: Overview of NIL

of the fact that the order of many tokens in a large-variance clone pair is preserved. In addition, large-variance clones share many consecutive tokens. Hence, to achieve scalable large-variance clone detection, the proposed technique reduces the number of code block comparisons by using an N-gram representation of token sequences and an inverted index. The proposed technique transforms code blocks in source code into token sequences in the *Preprocessing* phase and detects clones by comparing the token sequences in the *Clone detection* phase. In this study, we implemented the proposed technique as a tool, called NIL. NIL is written in the Kotlin language and currently targets only Java source code. The following subsections describe the *Preprocessing* and *Clone detection* phases.

3.1 Preprocessing

In the *Preprocessing* phase, NIL extracts code blocks from the target source code and transforms them into token sequences. NIL does not perform lexical analysis but simply divide each code block's text based on symbols (e.g., "+", "−", or braces), white spaces, or newlines, as done by SourcererCC. For example, when the code block shown in figure 1a is transformed into the token sequence, `protected`, `String`, `getPrompt`, `InputRequest`, `request`, ... With this transformation, lexical analyzers for other languages do not need to be implemented to extend NIL. The token sequence transformation is fast because lexical analysis is not necessary. In addition, NIL has a relatively low rate of false positives because it does not normalize identifiers, such as variable and function names. However, it may not detect clones whose identifiers have been changed (i.e., Type-2 clones). We discuss the impact of the lack of identifier normalization in Section 4.

Next, NIL generates N-grams from each token sequence. An N-gram is a chunk of consecutive N tokens. Figure 3 shows an example of generating 3-grams from the code block

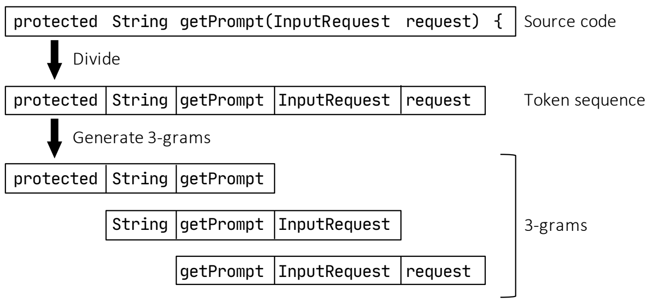

Fig. 3: Example of generating 3-grams

shown in figure 1a. In this example, three 3-grams are generated from the five tokens of on the first line in the code block. Even though large-variance clones include many modifications (statement insertions and deletions), many tokens other than the statements match consecutively (i.e., many N-grams match). Therefore, using N-grams is effective for scalable large-variance clone detection.

Then, NIL creates an inverted index from the generated N-grams. An inverted index is an information retrieval technology that allows the fast retrieval of documents that contains a word given as a query. It is often used in clone detection techniques [4]. NIL uses a dictionary whose keys are the hash values of N-grams, and values are the code blocks containing the corresponding N-gram as an inverted index. All code blocks containing an N-gram can be quickly obtained by looking up the hash value of the N-gram in the inverted index. Therefore, a pair of code blocks that share an N-gram (i.e., the pair is possibly a large-variance clone pair) can be obtained quickly using the inverted index.

3.2 Clone detection

After the *Preprocessing* phase, NIL performs *Clone detection* using the inverted index created in the *Preprocessing* phase and all code blocks. *Clone detection* is divided into three phases, namely location, filtration, and verification,

Algorithm 1: Clone Detection

```
Input: $C$ is a list of tokenized code blocks $\{c_1, c_2, \ldots, c_n\}$,
    Inverted Index $I$ of $C$, $N$ for size of N-gram, $\theta$ for filtering
    threshold, $\delta$ for verifying threshold
Output: All clone pairs $CP$
 1: $CP \leftarrow \phi$;
 2: for all each $c_i$ in $C$ do
 3:     // Location phase
 4:     // $CC$ represents clone candidates
 5:     $CC \leftarrow \phi$
 6:     for $j = 1, 2, \ldots, (c_i.len - N + 1)$ do
 7:         // $c_i[j]$ is $j$-th token in $c_i$'s token sequence
 8:         $n\_gram = concat(c_i[j], c_i[j+1], \ldots, c_i[j+N-1])$;
 9:         $key = hash(n\_gram)$;
10:         // $get$ is a function that returns values to which a given
            key is mapped in a given dictionary
11:         $CC = CC \cup get(I, key)$;
12:     end for
13:     // Filtration phase
14:     for all each $cc_j$ in $CC$ do
15:         /* $common\_ngrams$ is a function
            that computes the number of common N-grams between
            two given code blocks */
16:         $cn = common\_ngrams(c_i, cc_j)$;
17:         $m = min(c_i.len, cc_j.len)$;
18:         $filtration\_sim = cn/(m - N + 1)$;
19:         if $filtration\_sim < \theta$ then
20:             $CC = CC \setminus \{cc_j\}$;
21:         end if
22:     end for
23:     // Verification phase
24:     for all each $cc_j$ in $CC$ do
25:         /* $lcs$ is a function that computes the length of the LCS
            between token sequences of two given code blocks */
26:         $lcs\_len = lcs(c_i, cc_j)$;
27:         $verification\_sim = lcs\_len/min(c_i.len, cc_j.len)$;
28:         if $verification\_sim \geq \delta$ then
29:             $CP = CP \cup (c_i, cc_j)$;
30:         end if
31:     end for
32: end for
33: return $CP$;
```

as done by LVMapper [2]. First, NIL selects a code block from all code blocks prepared in the *Preprocessing* phase as the target code block. Then, in the location and filtration phases, NIL identifies the clone candidates of the target code block using an N-gram and the inverted index. Next, in the verification phase, NIL verifies that the target code block and the clone candidates are clone pairs by calculating the LCS. These phases are performed for each code block to detect all clone pairs in the target source code. Algorithm 1 shows the *Clone detection* algorithm. The three phases of *Clone detection* are described in detail below.

3.2.1 Locaion phase

In the location phase, NIL collects the clone candidates of the target code block using the inverted index. Lines 3–11 in Algorithm 1 are the location phase.

First, NIL generates N-grams from the token sequence of the target code block. M–N+1 N-grams are generated from a token sequence with length M. Next, a hash value is calculated for each N-gram. This hash value is used as a query when looking up values in the inverted index. Finally, NIL applies the hash values to the inverted index and collects code blocks that contain the N-gram whose hash value is the same as the given hash value. The obtained code blocks are referred to as the clone candidates of the target code block.

3.2.2 Filtration phase

In the filtration phase, NIL removes code blocks that unlikely to be clones from the clone candidates collected in the location phase. Lines 13–22 in Algorithm 1 are the filtration phase. It is necessary to reduce the number of clone candidates for scalable and fast clone detection because NIL performs the LCS calculation, which is a time-consuming process, in the verification phase. NIL filters clone candidates based on a feature of large-variance clones.

As described in Section 3.1, the two code blocks of a large-variance clone pair share a certain number of N-grams. Hence, if two code blocks share few N-grams, the pair is unlikely to be a large-variance clone pair. Based on this feature, NIL calculates $filtration\_sim$, defined below, between the target code block and each clone candidate.

$$filtration\_sim(c_1, c_2) = \frac{common\_ngrams(c_1, c_2)}{min(ngrams(c_1), ngrams(c_2))}$$

$$common\_ngrams(c_1, c_2) = |ngrams(c_1) \cap ngrams(c_2)|$$

where c_1 and c_2 are two code blocks with lengths $|c_1|$ and $|c_2|$, respectively. $ngrams(c_1)$ and $ngrams(c_2)$ are the numbers of N-grams generated from code blocks c_1 and c_2, respectively. Because of the large number of statement insertions and deletions in large-variance clones, the two code blocks may have significantly different token sequence lengths. We use min in the denominator so that $filtration\_sim$ can be properly calculated even such cases. NIL removes clone candidates whose $filtration\_sim$ is less than filtration threshold θ.

3.2.3 Verification phase

In the verification phase, NIL checks whether the target code block and each clone candidate are a true large-variance clone pair. Lines 23–31 in Algorithm 1 are the verification phase. As mentioned in Section 3.2.2, one of the features of large-variance clones is that the common subsequence between token sequences of large-variance clones is long even if there are a large number of insertions and deletions. Therefore, NIL calculates the LCS between the target code block and each clone candidate and measures the similarity of the pair based on the length of the LCS. The similarity function $verification\_sim(c_1, c_2)$ is expressed as following

$$verification\_sim(c_1, c_2) = \frac{lcs(c_1, c_2)}{min(|c_1|, |c_2|)}$$

where c_1 and c_2 are token sequences with lengths $|c_1|$ and $|c_2|$, respectively, and $lcs(c_1, c_2)$ is the length of the LCS between c_1 and c_2.

We use *min* as the denominator of the similarity function to detect large-variance clones, as done in the studies on CCAligner [5] and LVMapper [2] even if the lengths of the token sequences differ greatly.

4. Evaluation

We evaluated NIL in terms of

- large-variance clone detection accuracy,
- general clone detection accuracy, and
- scalability.

In the following subsections, we first optimize the N-gram size based on a balance between recall and execution time. Next, we evaluate large-variance clone detection accuracy in terms of precision and recall. Then, we evaluate general clone detection using two commonly used benchmarks. Finally, we evaluate scalability by measuring execution time for various input sizes. Additionally, we compare the above results to those for four state-of-the-art tools [2,3,5,6]. Table 1 shows these clone detectors and their settings. These settings were taken from the prior studies [2,3,5]. Note that the threshold δ for verification of LVMapper is variable and that δ takes the following values depending on the number of the lines of clone l.

$$\delta = \begin{cases} 0.7 & \text{if } 6 < l \leq 10, \\ 1 - 0.03 \times l & \text{if } 10 < l \leq 20, \\ 0.4 & \text{if } 20 < l \end{cases}$$

4.1 Summary

First, we summarize the results of this evaluation. We found that NIL has a high precision of 87% and a high recall of 100% in large-variance clone detection. These values are the highest among the tested large-variance clone detectors [2,5]. In general Type-1, Type-2, and Type-3 clone detection, NIL's accuracy is equivalent to that of the existing clone detectors, including large-variance clone detectors, and its precision is higher than that of large-variance clone detectors. Moreover, we confirmed that NIL is the fastest at detecting clones in large codebases (1–250 MLOC) among the tested clone detectors.

4.2 Parameter setting

NIL requires three parameters, namely N for N-grams, filtration threshold θ, and verification threshold δ. We set δ to 0.7, which is often used in clone detectors [3,6]. We set θ to 0.1, as done for LVMapper. N must be carefully selected because it has a large impact on performance (e.g., execution time and clone detection accuracy). If N is set to a too small value, the recall of clone detection will increase because more code blocks will share the same N-grams. However, because more code blocks are identified as clone candidates in the location phase, the number of comparison targets in the filtration and verification phases increases, resulting in a longer execution time. Therefore, to optimize the N, we executed NIL with $N = 1$–9 and measured the execution time and clone detection recall for each N value. We used BigCloneEval [7] to measure recall. BigCloneEval automatically measures the recall of clone detectors using BigCloneBench [8].

Table 1: Settings for various clone detectors

Tool	Settings
LVMapper	Min length 6 lines, window size $k = 3$, filtering threshold $\theta = 0.1$, verification threshold δ is dynamic.
CCAligner	Min length 6 lines, window size $q = 6$, edit distance $e = 1$, min 60% similarity.
SourcererCC	Min length 6 lines, min 70% similarity.
NiCad	Min length 6 lines, max length 20,000 lines, blind renaming, identifier abstraction, min 70% similarity.

Table 2 shows the results for each N value. For $N < 5$, an increase in N significantly reduces the execution time without lowering recall by more than one point compared to when $N = 1$. For $N = 5$ and 6, execution time decreases but recall also decreases. For $N > 6$, execution time does not significantly decrease. Therefore, considering the balance between execution time and recall, $N = 5$ is the optimal value.

4.3 Large-variance clone detection

We evaluated the large-variance clone detection accuracy of NIL in terms of precision and recall and compared the results to those for existing large-variance clone detectors, namely CCAligner and LVMapper.

4.3.1 Precision

Precision is the ratio of correct clones detected to all clones detected. A clone detector with higher precision provides more accurate results. In general, precision is measured via a manual validation of the clones detected by the target tool. In this study, we used Ant and Maven, which were used in the prior studies on CCAligner [5] and LVMapper [2] to measure precision. JDK1.2.2 and OpenNLP, also used in the above studies, were not used here because we could not find the source code for JDK1.2.2, which has been end-of-lifed and we considered OpenNLP to be unsuitable for manual validation because it is a machine learning library whose source code is difficult to read (e.g., it includes repetitive array manipulation code). We used the following procedure to measure the precision of each tool:

(1) we input the target source code into each tool,
(2) we randomly selected 100 large-variance clone pairs with more than 10 lines for each tool and target system, and
(3) we manually confirmed whether the clone pairs were correct.

To remove bias in the manual validation, the detected large-variance clone pairs of a given tool were validated without knowledge of the tool used for detection. Table 3 shows the number of files and total LOC for Ant and Maven.

Table 4 shows the number of large-variance clone pairs detected by each tool and the precision for each tool[*2]. The number of large-variance clones detected by NIL detected was almost same as that of LVMapper and more than that of CCAligner. The manual validation results indicate that NIL had a high average precision of 87% whereas LVMapper and CCAligner had low average precision values of about 60% and 40%, respectively. We considered this difference in precision to be due to LVMapper and CCAligner normalizing the identifiers in code blocks. After checking the large-variance clones detected by LVMapper and CCAligner, we found that code blocks with consecutive assignment statements, such as constructors, and consecutive `if`-statements were incorrectly detected as large-variance clones. In contrast, NIL detected large-variance clones more precisely because it does not perform identifier normalization.

In addition, NIL had a higher precision than that of LVMapper and detected a similar number of large-variance clone pairs, indicating that it can detect many large-variance clones that LVMapper cannot. After checking large-variance clones that NIL detected but LVMapper did not, we found that the large-variance clone pairs share a small number of consecutive lines. LVMapper regards three consecutive lines in code blocks as code windows (see Section 2) and identifies clone candidates based on these code widows. Therefore, if a pair of code blocks shares a little or no code window, LVMapper cannot detect the pair as a large-variance clone pair. In contrast, NIL can detect such large-variance clones because it uses an N-gram representation of token sequences.

Table 3: Target systems

System	# Files	LOC
Ant 1.10.1	895	109,073
Maven 3.5.0	698	60,471

Table 4: Large-variance clone detection results

Tool	System	# Large-variance clones	Precision (%)
NIL	Ant	354	86.0
	Maven	398	88.0
LVMapper	Ant	355	64.0
	Maven	389	60.0
CCAligner	Ant	184	43.0
	Maven	284	40.0

[*2] Manual validation descriptions are available at `https://zenodo.org/record/4490845`

4.3.2 Recall

Recall is the ratio of clones detected by a tool to the total number of true clones in the target codebase. A clone detector with a higher recall can detect true clones more exhaustively. To evaluate recall, all true clone pairs in the target codebase are required. However, it is not realistic to manually check all code block pairs in a system to determine the total number of true clone pairs.

Therefore, we generated large-variance clone pairs automatically using mutation techniques to evaluate recall. Mutation techniques are frequently used for clone detector recall evaluation [9]. The studies on CCAligner [5] and LVMapper [10] used them to evaluate large-variance clone detection recall. In this study, we reproduced large-variance clones by randomly inserting various numbers of statements into original code blocks (i.e., the large-variance clones are mutants of the original code blocks). We targeted JDK6 and Apache Commons[*3] because they were used in the study on CCAligner [5]. We randomly selected 100 functions with 15–50 lines in these systems as the original code blocks. A minimum length 15 lines is often used for recall evaluation using mutation techniques [3,9]. We used a maximum length of 50 lines because if the number of lines of the original code blocks is too large, even if a large number of statements are inserted into the code blocks, the generated clones will not be large-variance clones. For example, if 20 lines of statements are inserted into an original code block, if the number of lines of a code block is 100, the ratio of lines of the clone pair is $100/120 > 0.7$, and thus the clone pair does not satisfy the large-variance clone definition. To reproduce large-variance clone pairs, 1–20 one-line statements were inserted into each code block at random locations. 20 clone pairs were generated per original code block, for a total of 2,000 clone pairs, including large-variance clone pairs[*4].

Figure 4 shows the recall of NIL, LVMapper, and CCAligner for clone pairs generated by inserting various numbers of statements. NIL detected all generated clone pairs. This is because even though many statements are inserted into a code block, the order of many tokens between the large-variance clone pair is preserved, and thus the clone pair shares a certain number of N-grams and has a long common subsequence. In summary, using N-gram-based clone candidate identification and token-LCS-based clone validation is effective in large-variance clone detection.

[*3] A project that provides open-source reusable Java components
[*4] The generated large-variance clone pairs are available at `https://zenodo.org/record/4491016`

Table 2: Recall and execution time results for each N value

N	1	2	3	4	5	6	7	8	9
Type-2	97.5	97.5	97.5	97.5	96.6	96.3	96.1	95.8	93.3
Very Strongly Type-3	93.5	93.5	93.5	93.5	93.5	93.1	92.9	92.7	92.5
Strongly Type-3	68.4	68.4	68.4	68.3	67.1	65.3	64.6	63.7	62.0
Moderately Type-3	11.2	11.2	11.2	11.1	10.6	9.9	9.3	8.7	7.8
Execution time	> 24h	2h 13m 52s	21m 42s	5m 56s	2m 58s	1m 13s	1m 4s	1m 2s	57s

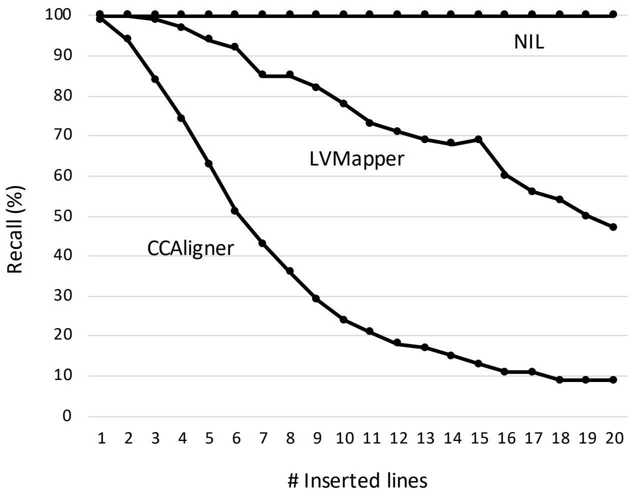

Fig. 4: Recall results for various numbers of inserted lines

On the other hand, as shown in Figure 4, the recall of LVMapper and CCAligner decreased with increasing number of inserted lines. LVMapper can also detect clones in which a large number of statements are inserted because it verifies clones based on line-based LCS. However, our evaluation results show that the recall of LVMapper decreased with increasing number of inserted lines. This is because LVMapper failed to identify many large-variance clones in its locating phase. LVMapper identifies a pair of code blocks that share some code windows (see Section 2) as a clone pair. Therefore, with increasing number of inserted lines, the number of shared code windows decreases, and thus LVMapper failed to identify a pair of code blocks as a large-variance clone pair. CCAligner uses e-mismatch code windows to identify clone candidates, and this affects its recall. In addition, CCAligner uses code windows in verification and thus fails to detect large-variance clones.

4.4 General clone detection

We evaluated the general Type-1, Type-2, and Type-3 clone detection accuracy of NIL using two benchmarks, namely Mutation Framework [9] and BigCloneEval [7]. In addition, we compared the results of NIL to those of existing state-of-the-art tools, namely CCAligner, LVMapper, SourcererCC, and NiCad.

4.4.1 Mutation Framework

Mutation Framework automatically generates clone pairs based on mutation techniques. We executed Mutation Framework with all the default settings and input the generated clones[*5] into each clone detector. Table 5 shows the results of recall for each tool measured by Mutation Framework. NIL detected all clone pairs generated by Mutation Framework.

4.4.2 BigCloneEval

BigCloneEval [7] automatically measures the recall of clone detectors using BigCloneBench [8]. We also measured precision, as done in the prior studies [2,3,5]. For each tool, we randomly selected 400 of the detected clone pairs from BigCloneBench and validated them manually. The clones were shuffled, and the validation was conducted without knowledge of the clone source.

Table 6 shows the results of recall[*6] and precision for each tool on BigCloneBench[*7]. As shown, NIL has a high recall of 96% for Type-2 clone detection even though it does not normalize identifiers in the *Preprocessing* phase. NIL had the second highest recall of Moderately Type-3 clones, which contain large-variance clones, behind only LVMapper. We considered that this is because normalizing identifiers is necessary to detect most Moderately Type-3 clones.

The precision (see bottom of Table 6) of both LVMapper and CCAligner, which are large-variance clone detectors, was low. In contrast, that of NIL was very high (94%). Even though SourcererCC and NiCad also had high precision, they had poor large-variance clone detection performance. Therefore, compared to the existing tools, NIL has equivalent general clone detection accuracy and higher precision than that of the existing large-variance clone detectors.

4.5 Scalability

We evaluated the scalability of NIL using codebases with various sizes and compared the execution time of NIL to those of the existing tools. We used IJaDataset [12], a large inter-project Java dataset, as done in the prior studies [2,3,5]. We created datasets with 1, 10, 100, and 250 MLOC[*8]. We used CLOC [13] to measure the LOC of the datasets. A computer with a quad-core CPU and 12 GB of memory was used for the evaluation, as done in the prior studies [3,5].

Table 7 shows the execution times for each tool for various input sizes. As shown, the execution time of NIL is the shortest for all input sizes. Even though both LVMapper and SourcererCC detected clones from the 250-MLOC codebases, their execution times are longer than three days, indicating poor scalability. In addition, CCAligner and NiCad was not able to complete detecting clones from the 100- and 10-MLOC codebases, respectively, indicating their limited scalability. Therefore, NIL has the highest scalability.

Table 5: Recall results for Mutation Framework

Tool	NIL	LVMapper	CCAligner	SourcererCC	NiCad
Type-1	100.0	100.0	100.0	100.0	100.0
Type-2	100.0	100.0	100.0	100.0	100.0
Type-3	100.0	99.9	99.9	100.0	100.0

[*5] The generated clone pairs are available at `https://zenodo.org/record/4491052`

[*6] Note that in our experiments, BigCloneEval reported different recall of the existing clone detectors from the prior studies. Type-1 recall of some clone detectors was 99.9% because BigCloneBench contains faulty clone pairs [11].

[*7] Manual validation descriptions are available at `https://zenodo.org/record/4493069`

[*8] These datasets and an executable file of NIL are available at `https://zenodo.org/record/4491208`.

5. Threats to validity

To measure precision for each tool, we manually validated clones each tool detected, as done in the prior studies [2,3,5]. Because the clone detector names were not disclosed during the manual validation, there was no bias in the evaluation. However, because the criteria for whether a pair of code fragments is a clone pair can vary, manual validation by other researchers may yield different values. To ensure the validity of this study, the clones used in the manual validation are made public so that a third party can conduct replication studies.

In this study, we used the widely used benchmarks, BigCloneEval [7] and Mutation Framework [9] to evaluate the recall of clone detectors. Different results may be obtained using other benchmarks.

In this study, we targeted only the Java language. Different results may be obtained for other languages.

It is known that the accuracy and execution time of a clone detector is greatly influenced by its settings [14]. In this study, we optimized the N value for N-grams. However, the filtration threshold θ and the verification threshold δ were set based on values used by other clone detectors. The results can be improved by optimizing these values for NIL.

6. Conclusion

In this study, we proposed a clone detection technique for the scalable detection of large-variance clones from a large codebase and described its implementation, called NIL. NIL uses N-grams, an inverted index, and the LCS to detect large-variance clones. Our experimental results show that NIL has higher precision and recall in large-variance clone detection than those of existing large-variance clone detectors. In addition, the general Type-1, Type-2, and Type-3 clone detection accuracy of NIL is equivalent to that of existing state-of-the-art tools. Moreover, NIL can detect clones from large codebases more quickly than do existing clone detectors.

As future works, we consider doing research on software engineering applications such as code recommendation and completion, refactoring, and bug propagation for large-variance clones using NIL.

Acknowledgments This work was supported by MEXT/JSPS KAKENHI 20H04166.

References

[1] Bellon, S., Koschke, R., Antoniol, G., Krinke, J. and Merlo, E.: Comparison and Evaluation of Clone Detection Tools, *IEEE Transactions on software engineering*, Vol. 33, No. 9, pp. 577–591 (2007).

[2] Wu, M., Wang, P., Yin, K., Cheng, H., Xu, Y. and Roy, C. K.: LVMapper: A Large-Variance Clone Detector Using Sequencing Alignment Approach, *IEEE Access*, Vol. 8, pp. 27986–27997 (2020).

[3] Sajnani, H., Saini, V., Svajlenko, J., Roy, C. K. and Lopes, C. V.: SourcererCC: Scaling Code Clone Detection to Big-Code, *Proceedings of the 38th International Conference on Software Engineering*, pp. 1157–1168 (2016).

[4] Hummel, B., Juergens, E., Heinemann, L. and Conradt, M.: Index-based code clone detection: incremental, distributed, scalable, *Proceedings of 2010 IEEE International Conference on Software Maintenance*, pp. 1–9 (2010).

[5] Wang, P., Svajlenko, J., Wu, Y., Xu, Y. and Roy, C. K.: CCAligner: a token based large-gap clone detector, *Proceedings of the 40th International Conference on Software Engineering*, pp. 1066–1077 (2018).

[6] Cordy, J. R. and Roy, C. K.: The NiCad Clone Detector, *Proceedings of 2011 IEEE 19th International Conference on Program Comprehension*, pp. 219–220 (2011).

[7] Svajlenko, J. and Roy, C. K.: BigCloneEval: A Clone Detection Tool Evaluation Framework with BigCloneBench, *Proceedings of 2016 IEEE International Conference on Software Maintenance and Evolution*, pp. 596–600 (2016).

[8] Svajlenko, J. and Roy, C. K.: Evaluating Clone Detection Tools with BigCloneBench, *Proceedings of 2015 IEEE International Conference on Software Maintenance and Evolution*, pp. 131–140 (2015).

[9] Svajlenko, J. and Roy, C.: The Mutation and Injection Framework: Evaluating Clone Detection Tools with Mutation Analysis, *IEEE Transactions on Software Engineering* (2019).

[10] Wu, M., Wang, P., Yin, K., Cheng, H., Xu, Y. and Roy, C. K.: LVMapper: A Large-variance Clone Detector Using Sequencing Alignment Approach (2019).

[11] Inoue, K., Miyamoto, Y., German, D. M. and Ishio, T.: Code Clone Matching: A Practical and Effective Approach to Find Code Snippets, *arXiv preprint arXiv:2003.05615* (2020).

[12] [Online]: Ambient Software Evolution Group: IJaDataset 2.0., `http://secold.org/projects/seclone` (2021).

[13] [Online]: CLOC: Count lines of code, `http://cloc.sourceforge.net/` (2021).

[14] Wang, T., Harman, M., Jia, Y. and Krinke, J.: Searching for Better Configurations: A Rigorous Approach to Clone Evaluation, *Proceedings of the 2013 9th Joint Meeting on Foundations of Software Engineering*, pp. 455–465 (2013).

Table 6: Recall and precision results for BigCloneBench

Tool	NIL	LVMapper	CCAligner	SourcererCC	NiCad
Type-1 Recall	99.9	99.9	99.8	93.8	99.9
Type-2 Recall	96.6	99.2	98.9	96.6	99.2
Very Strongly Type-3 Recall	93.5	98.1	97.4	68.2	98.4
Strongly Type-3 Recall	67.1	81.8	69.0	59.0	69.7
Moderately Type-3 Recall	10.6	19.1	10.0	4.8	0.5
Precision	94.0	58.5	33.7	99.2	80.2

Table 7: Scalability results

Tool	NIL	LVMapper	CCAligner	SourcererCC	NiCad
1 M	10s	29s	52s	3m 1s	1m 48s
10 M	1m 38s	13m 38s	26m 3s	27m 37s	—
100 M	1h 38m 29s	17h 23m 39s	—	19h 38m 5s	—
250 M	7h 40m 7s	3d 13h 47m 39s	—	5d 6h 55m 1s	—