UNIVERSITY
OF OSLO

Master's thesis

Incremental clone detection for IDEs using dynamic suffix arrays

CCDetect-LSP: Detecting duplicate code incrementally in IDEs

Jakob Konrad Hansen

60 ECTS study points

Department of Informatics
Faculty of Mathematics and Natural Sciences

Spring 2023

Jakob Konrad Hansen

Incremental clone detection for IDEs using dynamic suffix arrays

CCDetect-LSP: Detecting duplicate code incrementally in IDEs

Abstract

Duplicate code is present in most software today. The presence of duplicate code can negatively affect the maintainability and extensibility of the software, as modifying duplicated code can often lead to modifying every instance of the duplication. In this thesis we present CCDetect-LSP, a duplicate code detection tool which targets the IDE scenario and is both programming language- and IDE agnostic. Duplicate code detection can be expensive in terms of time and memory, which means most detection algorithms cannot be continuously run in a live environment, such as while editing code in an IDE. In order to facilitate live detection of duplicate code in a code base where small incremental edits are applied, our tool implements a novel detection algorithm which supports fast incremental updates. Our empirical results demonstrate that our incremental algorithm scales better than other non-incremental and incremental algorithms, when small edits are applied to a code base.

Acknowledgements

First and foremost, I would like to thank my supervisors Volker Stolz and Lars Tveito. Thank you Volker, for giving me countless advice, feedback and for allowing me to explore and pursue this topic, even if it was not what we originally thought this thesis would be about. Thank you Lars, for always being an incredibly positive and motivating supervisor, and for always being available for a coffee chat. This thesis would not be possible without them. I would also like to thank my family and my girlfriend, for always being there to support me during this process. Finally, I would like to thank my co-students and those who have taken the time to proofread this thesis.

Contents

1 Introduction **5**
- 1.1 Motivation and problem statement . . . 5
- 1.2 Our contribution . . . 6
- 1.3 Structure . . . 6

2 Background **7**
- 2.1 Software quality and duplicated code . . . 7
- 2.2 Code clones . . . 7
- 2.3 Code clone detection process and techniques . . . 9
- 2.4 Incremental clone detection . . . 11
- 2.5 Code clone IDE tooling . . . 12
- 2.6 The Language Server Protocol . . . 13
- 2.7 Preliminary algorithms and data structures . . . 13

3 Implementation: LSP server architecture **25**
- 3.1 Document index . . . 28
- 3.2 Displaying and interacting with clones . . . 29
- 3.3 Configuration . . . 29

4 Implementation: Initial detection **30**
- 4.1 Fragment selection . . . 31
- 4.2 Fingerprinting . . . 32
- 4.3 Suffix array construction . . . 33
- 4.4 Clone extraction . . . 37
- 4.5 Source-mapping . . . 38

5 Implementation: Incremental detection **43**
- 5.1 Affordable operations . . . 43
- 5.2 Updating the document index . . . 44
- 5.3 Updating fingerprints . . . 45
- 5.4 Computing edit operations . . . 45
- 5.5 Suffix array incremental updates . . . 53
- 5.6 Dynamic extended suffix arrays . . . 59
- 5.7 Dynamic LCP array updates . . . 61
- 5.8 Dynamic clone extraction and source-mapping . . . 65

6 Evaluation **67**
6.1 Verifying clones with BigCloneBench 67
6.2 Time complexity of detection 68
6.3 Benchmark performance 71
6.4 Memory usage 81
6.5 Languages and IDEs 82

7 Discussion **84**
7.1 Performance 84
7.2 Memory usage 86
7.3 Clones detected 87
7.4 IDE and language agnostic clone detection 88

8 Conclusion **90**
8.1 Related work 90
8.2 Future work 91
8.3 Conclusion 92

9 Appendix **98**

Chapter 1

Introduction

Duplicate code, often called code clones, is code which is more or less copied to different locations in a code base. Code clones occur in practically every large software project, and code clone analysis has therefore become a highly active field of research in the last decade. Many tools and algorithms have been developed to detect, manage and refactor code clones [23, p. 6]. Code clone detection in large code bases can be a time-consuming process. A consequence of this is that few existing tools are highly integrated into the workflow of software developers and are not designed to be run on a code base every time the code base changes. This thesis will therefore focus on efficient code clone detection and presents a novel algorithm and tool for detecting code clones in a real-time programming environment.

1.1 Motivation and problem statement

Many tools and algorithms exist for code clone detection. However, few of these have the capability of efficiently detecting clones in a real-time programming environment. A problem with existing algorithms is that most of them need to rerun the entire analysis whenever a change to a file happens. Incremental algorithms that do not recompute all clones from scratch are therefore interesting for use-cases such as while programming in integrated development environments (IDEs) and for analyzing different revisions of the same source code, but this type of algorithm has not been thoroughly explored for code clone analysis. Existing incremental algorithms either do not scale well in terms of time or memory, or are not designed to be run on a

Our proposed solution and tool addresses this issue by introducing a new algorithm that is capable of detecting and updating existing code clones as source code changes, which aims to be faster than redoing the analysis from scratch. CCDetect-LSP, the tool which implements this algorithm, is also programming language- and IDE agnostic, allowing programmers to seamlessly incorporate the detection of duplicated code into their existing development environment.

1.2 Our contribution

CCDetect-LSP provides exact code clone detection capabilities in a real-time IDE environment. The tool allows the user to list and interact with all the clones in the code base, jump between matching clones, and get fast feedback while editing code in order to determine which clones are introduced or eliminated.

Existing incremental clone detection tools either do not fit into an IDE scenario, are limited in terms of what clones they display, or have not been shown to scale well in terms of processing time or memory usage for larger codebases. Therefore, this thesis will focus on the following areas of code clone analysis.

Incremental clone detection: The main focus of this thesis is making the tool efficient in terms of incrementally updating the analysis whenever edits are performed in the IDE. Most clone detection algorithms either only list clones of a specific code snippet, or calculates clones from scratch in a manner which is too slow for an IDE scenario. Our algorithm is based on a novel application and extension of dynamic suffix arrays for clone detection, which can find clones and be efficiently updated. While suffix arrays have been used for clone detection before [28], we are not aware of any other attempts to use suffix arrays in an incremental setting. CCDetect-LSP can display all clones in the entire code base at once, and efficiently update the list of clones whenever a file is edited. Our results demonstrate that this algorithm scales better than a non-incremental suffix array algorithm and an existing incremental algorithm in terms of time, when small edits are applied to a code base.

IDE and language agnostic clone detection: CCDetect-LSP gives programmers the ability to view clones in their IDE. Utilizing features of the Language Server Protocol (LSP) such as diagnostics and code-actions [32], the tool provides clone analysis to any editor which implements LSP. As far as we are aware there are no other clone detection tools which utilizes LSP to provide clone analysis to IDEs. In addition, the tool is also language agnostic in the sense that it only needs a grammar for the parser generator Tree-sitter, for it to be analyzed [10].

1.3 Structure

Chapter 2 provides background on code clone analysis, existing tools and preliminary algorithms and data structures used in the implementation. Chapter 3-5 describes the implementation of CCDetect-LSP and the algorithms used for initial and incremental clone detection. In chapter 6, the tool is evaluated based on multiple criteria, and compared to other existing solutions. Chapter 7 discusses the results of the evaluation. Chapter 8 lists related and future work, and concludes the thesis.

Chapter 2

Background

2.1 Software quality and duplicated code

Software quality is hard to define. The term "quality" is ambiguous and is in the case of software quality, multidimensional. Quality in itself has been defined as "conformance to requirements" [14, p. 8]. In software, a simple measure of "conformance to requirements" is correctness, and a lack of bugs. However, software quality is often measured in other metrics, including metrics which are not directly related to functionality [26, p. 29]. Whenever duplicated code needs to be changed, it might require changes in multiple locations, which leads to metrics such as maintainability, analyzability and changeability being negatively affected. Studies generally report that software projects contain $10 - 15\%$ duplicated code, but some studies also report lower and higher percentages [2]. Even so, the percentage of duplication is generally considered to be of non-trivial size in many large code bases. Therefore, research into tools and techniques which can assist in reducing duplicated code will be of benefit to almost all software.

Duplicated code can lead to a plethora of anti-patterns in software. Anti-patterns are bad design decisions for software which can lead to technical debt. Technical debt occurs when developers make technical compromises that are beneficial in the short term, but increases complexity in the long-term [4, p. 111]. An example of this, in the context of duplicated code, is the "Shotgun-Surgery" anti-pattern [15, p. 66]. This anti-pattern occurs when a developer wants to implement a change, but needs to change code at multiple locations for the change to take effect. This is a typical situation which slows down development and reduces maintainability when the amount of duplicated code increases in a software project.

2.2 Code clones

Duplicated code is often described as "code clones". A pair of code snippets which are duplicated are considered clones of each other.

Definition 1 (Code snippet)**.** *A code snippet is a piece of contiguous source code in a larger software system.*

```
for (int i = 0; i < 10; i++) {
    print(i);
}
```

```
for (int i = 0; i < 10; i++) {
    // A comment

    print(i);
}
```

Figure 2.1: Type-1 clone pair

Definition 2 (Code clone)**.** *A code clone is a code snippet which is equal to or similar to another code snippet. The two code snippets are both code clones, and together they form a clone pair. Similarity is determined by some metric such as number of equal lines of code.*

Definition 3 (Clone class)**.** *A clone class is a set of code snippets where all snippets are considered clones of each other.*

The clone relation is a relation between code snippets which defines pairs of clones. The clone relation is reflexive and symmetric, but not necessarily transitive. The transitive property depends on the threshold for similarity when identifying code clones. Given

$$a \overset{clone}{\longleftrightarrow} b \overset{clone}{\longleftrightarrow} c$$

where a, b, c are code snippets and $\overset{clone}{\longleftrightarrow}$ gives the clone relation. a and c are both clones of b, but not necessarily similar enough to be clones of each other, depending on the threshold for similarity. If the threshold for similarity is defined such that only equal clones are considered clones, the relation becomes transitive, and clone classes are equal to the equivalence classes of the relation.

Code clones are generally classified into four types [23]. The types would classify two code snippets as code clones with an increasing amount of leniency. Therefore, type-1 code clones are very similar, while type-4 clones are not necessarily syntactically similar at all. When defining types, it is the syntactic and structural differences which is compared, not functionality. The set of code clones classified by a code clone type is also a subset of the next type, meaning all type-1 clones are also type-2 clones, but not vice versa.

Type-1 clones are syntactically identical. The only differences for a pair of type-1 clones are elements without meaning, like comments and white-space, which is often already ignored when source code is lexed and parsed. Figure 2.1 shows an example of a type-1 clone pair where only a comment and white-space is added to the snippet on the right.

Type-2 clones are structurally identical. Possible differences include changes to identifiers, literals and types. Type-2 clones are not much harder to detect than type-1 clones, since consistently renaming identifiers, literals and types allow a type-1 detection algorithm to find type-2 clones [6]. This type of clone is relevant to consider in refactoring scenarios when merging code clones as they can be relatively simple to parameterize in order to merge two clones with for example differing literals. The original locations of clones can then be replaced with a call to the merged code with different parameters. Figure 2.2 shows an example of a type-2 clone pair where the

```
for (int i = 0; i < 10; i++) {
    print(i);
}
```

```
for (int j = 5; j < 20; j++)
    {
    print(j);
}
```

Figure 2.2: Type-2 clone pair

```
for (int i = 0; i < 10; i++) {
    print(i);
}
```

```
for (int i = 0; i < 10; i++) {
    print(i);
    print(i*2);
}
```

Figure 2.3: Type-3 clone pair

numbers in the for loop initialization and condition could be parameterized to correctly merge the two clones.

Type-3 clones are required to be structurally similar, but not equal. Differences include statements that are added, removed or modified. This clone type relies on a threshold θ which determines how structurally different snippets can be in order to be considered type-3 clones [23, p. 6]. The granularity for this difference could for example be based on the number of differing tokens or lines. Detecting this type of clone is generally computationally harder than detecting type-1 and type-2 clones. Figure 2.3 shows two code snippets where the code snippet on the right has an added statement. In this example there is a one line difference between the two snippets, so if the similarity is based on differing lines and $\theta \geq 1$, the two snippets would be considered Type-3 clones.

Type-4 clones have no requirement for syntactical or structural similarity, but are generally only relevant to detect when they have similar functionality. Detecting this type of clone is very challenging, but attempts have been made using program dependency graphs [47]. Figure 2.4 shows two code snippets which have no clear syntactic or structural similarity, but is functionally equal.

Type-1 clones are often referred to as "exact" clones, while Type-2 and Type-3 clones are referred to as "near-miss" clones [49, p. 1].

2.3 Code clone detection process and techniques

The Code clone detection process is generally split into (but is not limited to) a sequence of steps to identify clones [23]. This process is often a pipeline of input-processing steps before finally comparing fragments against each other and filtering. The steps are generally as follows:

1. **Pre-processing**: Filter uninteresting code that we do not want to check for clones, for example generated code. Then partition code into a set of fragments, depending on gran-

```
print((n*(n-1))/2)
```

```
int sum = 0;
for (int i = 0; i < n; i++) {
    for (int j = i+1; j < n;
        j++) {
        sum++;
    }
}
print(sum);
```

Figure 2.4: Type-4 clone pair

ularity such as methods, files or lines.

2. **Transformation**: Transform fragments into an intermediate representation, with a source-map which gives the location to the original code. An algorithm could potentially do multiple transformation before arriving at the wanted representation

   (a) Extraction: Transform source code into the input for the comparison algorithm. Can be tokens, AST, dependency graphs, suffix tree, etc.

   (b) Normalization: Optional step which removes superficial differences such as comments, whitespace and identifier names. Often useful for detecting type-2 clones.

3. **Match detection**: Perform comparisons which outputs a set of candidate clone pairs.

4. **Source-mapping / Formatting**: Convert candidate clone pairs from the transformed code back to clone pairs in the original source code.

5. **Post-processing / Filtering**: Ranking and filtering manually or with automated heuristics

6. **Aggregation**: Optionally aggregating sets of clone pairs into clone sets

Matching techniques are the techniques used in the matching stage of the algorithm which pairs fragments as clone-pairs. Most matching technique will also require specific pre-processing to be done in the earlier steps, for example building an AST. Some of the most explored techniques are as follows [37]:

Text-based approaches do little processing of the source code before comparing. Simple techniques such as fingerprinting or incremental hashing have been used in this approach.

Token-based approaches transform source code into a stream of tokens, similar to lexical scanning in compilers. The token stream is then scanned for duplicated subsequences of tokens. Since lexers often filter out superficial differences such as whitespace, indentation and comments, this approach is more robust to such differences. Concrete names of identifiers and values can be abstracted away when comparing the token-stream, therefore type-2 clones can easily be identified. Type-3 clones can also be identified by comparing the fragments tokens and keeping

clone pairs with a lexical difference lower than a given threshold. This can be solved with dynamic programming [7]. A common approach to detect clones using token-streams is with a suffix-tree [6]. A suffix-tree can solve the *Find all maximal repeats* problem efficiently, which essentially is the same problem as finding clone pairs [19, p. 143]. A suffix array can also be used instead, which requires less memory [1]. This type of code clone detection is very fast compared to more intricate types of matching techniques.

Syntactic approaches transform source code into either concrete syntax trees or abstract syntax trees and find clones using either tree matching algorithms or structural metrics. For tree matching, the common approach is to find similar subtrees, which are then deemed as clone pairs. One way of finding similar subtrees is to compare subtrees with a tolerant tree matching algorithm for detecting type-3 clones [9]. Variable names, literal values and types may be abstracted to find type-2 clones more easily. Metrics-based techniques gather metrics for code fragments in the tree and uses the metrics to determine if the fragments are clones or not. One way is to use fingerprint functions where the fingerprint includes certain metrics, and compare the fingerprints of all fragments to find clones [25].

Hybrid approaches combine multiple approaches in order to improve detection capabilities. For example a common approach which Zibran et al. [49] has implemented is a hybrid algorithm combining both token-based suffix trees for type-1 and type-2 clone detection, with the dynamic programming algorithm [7] for type-3 clone detection. Another example of a hybrid approach is the tool named Siamese [35], which uses multiple intermediate representations to get a better accuracy for clone detection.

2.4 Incremental clone detection

Incremental clone detection avoids computation of already computed results when performing code clone detection on consecutive revisions of a code base. A revision of a code base is one version of a code base which has been changed from a previous revision, and once edited, becomes the next revision. A revision of the code base could for example be a certain Git commit of the code base, or one could say a new revision of the code base is created for each edit performed while programming. These two scenarios have respectively been called the evolution scenario and the IDE scenario [17]. Incremental clone detection could be realized in the form of either storing clones between each revision, or by maintaining a data structure which is fast to update when the input updates, and also facilitates fast extraction of clones. Since in most cases, an entire code base will not change drastically in consecutive revisions, many code clones and the general structure of most of the source code will remain, and therefore a lot of processing can be avoided. However, this is not a simple problem, since changes in a single location can affect clone detection results across the entire codebase.

In order to incrementally detect code clones, an initial algorithm computes the initial clones, and for successive revisions of the source code, this list is incrementally updated by an incremental algorithm. In some cases, this could be the same algorithm, but in general the incremental algorithm is faster than the initial algorithm, and maintains dynamic structures to perform efficient updates.

Göde and Koschke proposed the first incremental clone detection algorithm [17]. The algorithm employs a generalized suffix tree in which the amount of work of updating is only dependent

on the size of the edited code. This approach requires a substantial amount of memory, and is therefore limited in scalability.

Nguyen et al. [33] showed that an AST-based approach utilizing "Locality-Sensitive Hashing" can detect clones incrementally with high precision, and showed that incremental updates could be done in real-time (< 1 second) for source code with a size of 300KLOC (lines of code).

Hummel et al. [22] later introduced the first incremental and distributed clone detection technique for type-1 and type-2 clones. This approach utilizes a custom "clone index" data structure which can be updated efficiently. The implementation of this data structure is similar to that of an inverted index. This technique uses distributed computing to speed up its detection process.

More recently, Ragkhitwetsagul and Krinke [35] presented the tool Siamese, which as mentioned uses a novel approach of having multiple intermediate representations of source code to detect a high number of clones with support for incremental detection. The tool can detect up to type-3 clones, but will only return clones based on "queries" given to it by the user, and therefore does not allow listing all clones in the code base. Queries are either files or methods in source-code, which are then matched with other code clones.

2.5 Code clone IDE tooling

Developers are not always aware of the creation of clones in their code. *Clone-aware development* means including clone management as a part of the software development process [49]. Clone-aware development therefore requires programmers to be aware of and be able to identify code clones while programming. Since large software projects can contain a lot of duplication, it can be hard to keep track of and manage clones. Tools which help developers locate and deal with clones can be a solution to this. However, Rieger et al. claims that a problem with many detection tools is that the tools "report large amounts data that must be treated with little tool support" [36, p. 1]. Detecting and eliminating clones early in their lifecycle with IDE integrated tools could be a solution to the problem of dealing with too many clones.

There are multiple existing clone detection tools which integrates into IDEs. The IDE-based tools which exist can be categorized as follows [42, p. 8]:

- *Copy-paste-clones:* This category of tools deals only with code snippets which are copy-pasted from another location in code. These tools therefore only track clones which are created when copy-pasting, and does not use any other detection techniques. Therefore, this type of tool is not suitable for detecting clones which are made accidentally, since developers are aware that they are creating clones when pasting already existing code snippets.

- *Clone detection and visualization tools:* This category of tools has more sophisticated clone detection capabilities and will detect all code clones algorithmically, unlike copy-paste clone tools.

- *Versatile clone management:* This category of tools covers tools which provide more services than the above. Services like refactoring and simultaneous editing of clones fall under this category.

The following tools have been developed as IDE tools to allow for clone-aware development:

- Zibran et al. introduced a hybrid technique for performing real-time focused searches, i.e. searching for code clones of a selected code snippet [49]. This technique can also detect Type-3 clones. This algorithm was later implemented in the tool SimEclipse [42] which is a plugin for the Eclipse editor. This tool is also based on a suffix tree, but it is unclear if this suffix tree is updated incrementally.

- Another tool, SHINOBI, which is a plugin for the Visual Studio editor, can detect code clones in real-time without the need of the developer to select a code snippet [28]. However, the clones being displayed seem to only be clones of the source-code which is currently being edited. This is limiting if a developer is searching for locations of clones, as they cannot simply look up the location of all clones. SHINOBI can detect type-1 and type-2 code clones and uses a token-based suffix array approach to detect clones. The paper does not describe how the suffix array is updated, but as the paper was released before the first paper describing dynamic suffix arrays [39], one can assume that the suffix array is recomputed from scratch in each search.

- The modern IDE IntelliJ has a built-in duplication detection and refactoring service, it is able to detect type-1 and (some) type-2 code clones at a method granularity and can refactor to remove a clone-pair by replacing one of the clones with a method call to the other. The service does not seem to perform incremental updates.

2.6 The Language Server Protocol

Static analysis tools that integrate with IDEs are often tightly coupled to a specific IDE and its APIs, like parsing and refactoring support. This makes it difficult to utilize a tool in another IDE, since the APIs are no longer available. In order to make IDE-based static analysis tools more widely available, it is therefore useful for such tools to be made IDE agnostic. The Language Server Protocol (LSP) is a server-client protocol which attempts to solve this problem [32].

LSP is a protocol which specifies interaction between a client (IDE) and server (analysis tool) in order to provide language tooling for the client. The goal of the protocol is to avoid multiple implementations of the same language tools for every IDE and every language, allowing for IDE agnostic tooling. Servers which implement LSP will be able to offer IDEs code-completion, error-messages, go-to-definition and more. LSP also specifies generic code-actions and commands, which the LSP server communicates to the client in order to perform custom actions defined by the server.

Figure 2.5 shows a sample interaction between client and server using LSP. The client sends requests to a server in the form of JSON-RPC messages, and the server sends a corresponding response, also in the form of JSON-RPC messages.

2.7 Preliminary algorithms and data structures

Before our implementation is discussed, we will look at some preliminary algorithms and data structures which may provide useful insights in the following chapters.

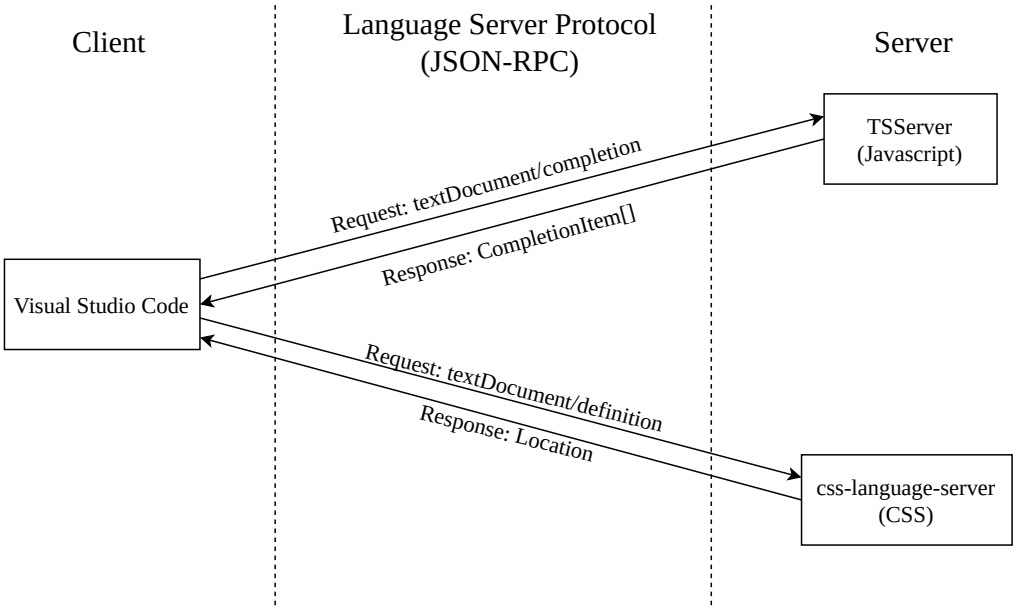


Figure 2.5: Example client-server interaction using LSP

Parsing and incremental-parsing

One of the first steps in most interpreters and compilers is to perform lexing and parsing on the source code. Lexing is the process of turning the source code into a set of tokens, turning strings such as “if”, “while”, “+” and “(” into singular values which are easier to work with. After the source code has been converted to a token-stream, the next step is to parse the source code into a tree structure called a syntax tree. A syntax tree is a tree which describes the structure of a program. For example an if-statement would usually be a node in the tree, with a condition and a list of statements as children. An abstract syntax tree (AST) is a tree where some details of the grammar may be left out in the tree structure, such as having concrete nodes of tokens such as “(” and “)”, which does not matter in a compiler or interpreters understanding of the program.

When parsing a program, the parser often relies on a grammar for the language, which specifies the structure of the syntax tree, and is usually given in the form of a BNF grammar. Given a BNF grammar, one could either write a parser which conforms to this grammar (often recursive-descent for LL(1) grammars), or use a parser generator, which generates the parser, given the grammar of the language. Parser generators can often generate parser which are more complex, which allows for more complex grammars, such as LALR(1) or LR(1) grammars.

A syntax tree is not only used for interpreting or compiling a language, it is also used to perform static analysis on the program. Static analysis is an analysis performed without running the program, and is often done in compilers in order to do for example type-checking or data-flow analysis. Syntax trees are in essence very useful to traverse or search a programs structure to look for certain properties of nodes or subtrees.

In our detection algorithm we will need to be able to parse our source code into an AST rep-

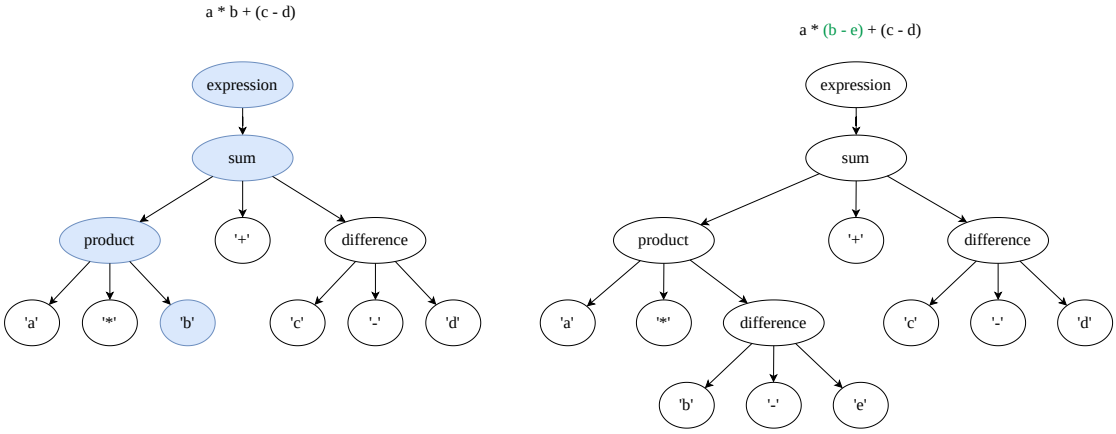


Figure 2.6: AST's of two expressions. Left AST could be incrementally parsed to the right AST if marked nodes are parsed again and the rest of the tree is reused.

resentation in order to select specific syntactic regions and extract the tokens. Therefore, we need a parsing algorithm. For an incremental scenario, a useful observation is that when a small edit is applied to a program, the syntax tree will usually not change too much. When source code is edited it would therefore be more efficient to use an incremental parsing algorithm which reuses parts of the old syntax tree, but updates it to reflect the new source code. Incremental parsing is the process of parsing only parts of a syntax tree whenever an edit is performed in a file, and reusing subtrees of the previous syntax tree which did not change. The motivation behind incremental parsing is to have a readily available syntax tree after every edit, while doing as little computing as possible to maintain it.

Tree-sitter is a parser generator tool which specializes in incremental parsing. It supports incremental parsing, error recovery and querying for specific nodes and subtrees [10]. These features combined makes Tree-sitter a powerful tool for editing and has been used for IDE features such as syntax-highlighting, refactoring and code navigation.

The incremental parsing algorithm in Tree-sitter[1] is inspired by and implemented similarly to Wagner's incremental parsing algorithm [46]. Each node in the AST is decorated with information on the range which it covers in the original source code. Whenever an edit is performed at location x in the source code, a new AST is built by first marking all the nodes where the range of the node contains the position x. If a node contains the position x and is marked, it needs to be parsed again, but subtrees rooted in nodes which are not marked can be reused, and can therefore be reused. Figure 2.6 shows on the left an expression and its corresponding AST. If the expression were to be changed to the expression on the right, the AST on the left could be incrementally parsed by parsing only the marked nodes, and reusing all the other nodes. Especially with the right-most difference node, we see how reusing an entire subtree in the AST could save time compared to normal parsing. In larger programs we can see significantly more reusage of subtrees, since an edit to a few lines would likely require few nodes to be parsed compared to the number of nodes reused.

[1]For a preliminary introduction to Tree-sitter incremental parsing, see the following conference talk:
`https://www.thestrangeloop.com/2018/tree-sitter---a-new-parsing-system-for-programming-tools.html`

Suffix trees

A classic algorithm for code clone detection traverses a suffix-tree in order to find maximal repeats in all suffixes of the input string S [49, 17]. While our algorithm is only related to suffix trees, this section gives some insight into the most common algorithm for clone detection.

A suffix of a string S is a substring $S[i..N]$, denoted where N is the length of S. A suffix of S is therefore any nonempty substring which reaches the end of S. The suffix at position i of S is denoted Suffix(S, i).

The suffix tree of a string S is a compressed trie where all the suffixes of S have been inserted into it. Consecutive nodes of the trie that only have one child are compressed into a single node. A common usage of a suffix tree is to determine whether a suffix exists in S, and where in S the suffix starts.

Figure 2.7 shows the suffix-tree for S = BANANA\$. \$ is often added as a unique terminal character, which simplifies some string problems. In order to determine where the suffix ANANA\$ exists in S, one can start from the root, and traverse the tree, choosing the child node which correspond to the next character of the suffix which has not been "matched" yet.

$$\text{root} \xrightarrow{A} \text{node} \xrightarrow{\text{NA}} \textit{node} \xrightarrow{\text{NA\$}} 1$$

Following this path, we see that the suffix ANANA\$ exists in S at index 1.

Suffix trees can be constructed in linear time with Ukkonen's algorithm which builds a larger and larger suffix tree by inserting characters one by one and utilizing some tricks to avoid inserting suffixes before it needs to, lowering the complexity [44].

This data structure also facilitates solving the maximal repeat problem. A repeat in a string S is a substring that occurs at least twice in S. A maximal repeat in S is a repeat which is not a substring of another repeat in S, meaning that the maximal repeat cannot be extended in any direction to form a bigger repeat. The problem of finding all maximal repeats can be solved with a suffix tree using the following theorem:

Theorem 1 (Repeats in suffix tree)**.** *Every internal node (except for the root) in a suffix tree corresponds to a substring which is repeated at least twice in T. The substring is found by concatenating the strings found on the path from the root of tree to the internal node.*

This theorem is explained by the fact that any internal node has at least two children, and a node having two children means that two suffixes share the same prefix up to that point. An algorithm which finds the maximal repeats would find the internal nodes which represents the longest strings.

The classic algorithm [49, 17] in terms of finding duplication in a string (such as source code) using suffix trees would find all repeats of length k, where k is the threshold for how long a clone needs to be. This can be found by traversing the suffix tree and looking at all internal nodes which represent a string of length $\geq k$. Every internal node which represents a string which is $\geq k$ would correspond to a substring of the source code which occurs at least twice. Finding where the duplication occurs can be done by finding all the leaves of the subtree rooted in the internal node, which each hold the position where the suffix starts in S. Since a substring

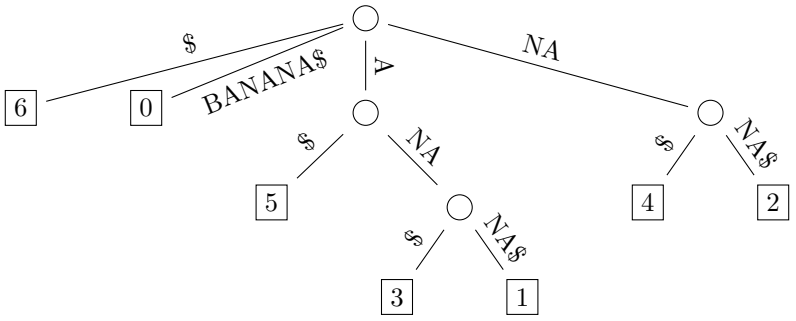


Figure 2.7: Suffix tree for S = BANANA$

can have multiple repeats of different lengths longer than k, different strategies can be used to select which substrings are selected or not, such as filtering out repeats which are not maximal or repeats which contain or overlap each other.

In figure 2.7, there are three repeats, ANA, A and NA. The only maximal repeat would be ANA, since A and NA are not maximal.

Suffix arrays

A problem with the suffix tree data structure is that the pointers and nodes require a lot of memory. The suffix array (SA) is a data structure which can achieve the same functionality, but with a memory efficient array instead. The suffix array of a string S contains a lexicographical sorting of all suffixes in S. The suffix array does not contain the actual suffixes, but it contains integers pointing to the index where the suffix starts in S. Conversely, the inverse suffix array (ISA) contains integers describing which rank the suffix at a given position has. The rank of a suffix is its lexicographical ordering in S. ISA is therefore the inverse array of SA, such that if $\text{SA}[i] = n$, then $\text{ISA}[n] = i$.

Definition 4 (Suffix array)**.** *Let S be a text of length N. The suffix array SA of S is an array of length N where $SA[i] = n$ if the suffix at $S[n..N]$ is the ith lexicographically smallest suffix in S.*

Definition 5 (Inverse suffix array)**.** *Let S be a text of length N. The inverse suffix array ISA of S is an array of length N where $ISA[i] = n$ if the suffix at $S[i..N]$ is the nth lexicographically smallest suffix in S.*

The Longest-common prefix (LCP) array describes how many common characters there are in a prefix between two suffixes which are next to each other in the suffix array. Since the suffix array represents suffixes in a sorted order, the prefix length between adjacent suffixes in SA will be the longest possible common prefix for each suffix. These are the values in the LCP array.

Definition 6 (LCP array)**.** *Let T be a text of length N and SA be the suffix array of T. The LCP array of T is an array of length N where LCP[i] = n if the suffix T[SA[i]..N] and T[SA[i-1]..N] has a maximal common prefix of length n. LCP[0] is undefined or 0.*

For example, in table 2.1, the LCP value at position 3 contains the number of characters in the longest-common prefix of suffixes starting at position 1 (ANANA$) and position 3 (ANA$).

Index	Suffix
0	BANANA$
1	ANANA$
2	NANA$
3	ANA$
4	NA$
5	A$
6	$

(a) Suffixes

Index	Suffix
6	$
5	A$
3	ANA$
1	ANANA$
0	BANANA$
4	NA$
2	NANA$

(b) Sorted suffixes

Index	SA	ISA	LCP
0	6	4	0
1	5	3	0
2	3	6	1
3	1	2	3
4	0	5	0
5	4	1	0
6	2	0	2

(c) SA, ISA and LCP

Table 2.1: T = BANANA$

These suffixes have 3 common characters in their prefix, therefore the LCP value at position 3 is 3.

Suffix arrays can be constructed in linear time in terms of the length of T. Many suffix array construction algorithms (SACA) have been discovered in the last decade [27], many of which run in linear time. An algorithm which has been shown to be very efficient in practice is Nong and Chan's algorithm based on recursive suffix sorting of smaller constructed strings, which will be used in our initial detection algorithm [34].

The extended or enhanced suffix array is a data structure storing both the suffix array and the additional LCP array, which has been shown to be as powerful as a suffix tree, in terms of what can be computed with it in the same time complexity [1]. An advantage of using an extended suffix array over a suffix tree is that it uses a smaller amount of memory, as suffix trees needs to store a lot of pointers and nodes, while the extended suffix array only need to store arrays of integers. The extended suffix array is the major data structure which will be used in our detection algorithm to find duplicate code, as explained in the upcoming implementation chapters.

Dynamic bitsets

For our detection algorithm, we need to utilize rank/select queries on strings in order to determine how the suffix array has changed. We will first discuss rank/select queries on bitsets, which will be used in the next section to extend this notion to strings as well.

A bitset is an array of bits, each bit representing either the value true or false. A bit with the value of 1 is usually referred to as a set bit, and a value of 0 is referred to as an unset or cleared bit. Operations on bitsets normally include at least an operation for setting the value at a position, and an operation for looking up the value at a position. Bitsets are useful for many problems, especially as a "succinct data structure". A succinct data structure is a data structure which attempts to use an amount of memory close to the theoretic lower bound, while still allowing effective queries on it. For example for a string of length n, we could store up to $O(n \log \sigma)$ bits before the bit vector exceeds the size of the string itself, where σ is the size of the string's alphabet.

The most well known queries to perform on bitsets is the rank/select queries.

Definition 7 (Rank query on bitset)**.** *A rank query $rank_1(i)$ on a bitset B, computes the number of set bits up to, but not including position i. Conversely, $rank_0(B)$ returns the number of unset*

bits up to position i.

Definition 8 (Select query on bitset)**.** *A select query $select_1(i)$ on a bitset B, computes the position of the ith set bit in B. Conversely, $select_0(i)$ returns the position of the ith unset bit in B.*

Jacobson's rank can perform rank/select queries on static bitsets in $O(1)$ time by pre-computing all answers in a space efficient table [24].

In some cases it is also useful to have a dynamic bitset, a bitset where insertions/deletions are supported as well.

Definition 9 (Dynamic bitset)**.** *A dynamic bitset is a bitset which in addition to other operations allow inserting and deleting bits.*

An insert operation $insert(i, v)$ on a bitset B inserts the value v at position i in B, pushing all values at position $\geq i$ one position up.

A delete operation $delete(i)$ on a bitset B removes the value at position i in B, pushing all values at position $> i$ one position down.

A simple implementation of a dynamic bitset could implement the whole bitset as a single array of bytes B, which allows for accessing values in $O(1)$ time, but insertions and deletions take $O(n)$ time, where n is the number of bits in B.

Another implementation which speeds up the insertions and deletions represents the bitset as a balanced tree containing multiple smaller bitsets. To represent a bitset of n bits, we can divide the bits into smaller bitsets, such that $O(\frac{n}{\log(n)})$ bitsets each contains $O(log(n))$ bits. Since the bitsets are now of size $O(log(n))$, insertions and deletions take only $O(log(n))$ time. The bitsets reside at the leaves of our balanced tree, and internal nodes contain only two integers, storing the number of bits in the left subtree (N), and the number of set bits in the left subtree (S). To access, insert or delete on index i, the tree is traversed to find the correct bitset where the ith bit is located, where the operation is done at that position. Finding the correct bitset and position is done by utilizing N and S in each node which is traversed. All operations now run in $O(\log(n))$ time, since traversing the tree to the correct bitset takes $O(\log(\frac{n}{\log(n)}))$ and performing the operation takes $O(\log(n))$ time. Keeping the tree balanced is done by splitting a leaf-node bitset into two nodes when the leaf-node bitset has reached a certain size (such as $2 \times \log(n)$) and rebalancing the tree after the split. Similarly, when considering deletions, two leaf-node bitsets can be merged to their parent node when their combined number of bits has decreased to $\log(n)$.

Figure 2.8 shows how a dynamic bitset tree is structured, and Algorithm 1 shows how to access a value in the tree. Traversing the tree to compute rank/select queries can be done similarly by summing set bits in left-subtrees (rank) or selecting which subtree to descend based on the number of set bits (select).

Wavelet tree and wavelet matrix

We can extend the notion of rank/select queries to strings as well.

Definition 10 (Rank query on string)**.** *A rank query $rank_c(i)$ on a string S computes the number of occurrences of a character c in S up to, but not including position i.*

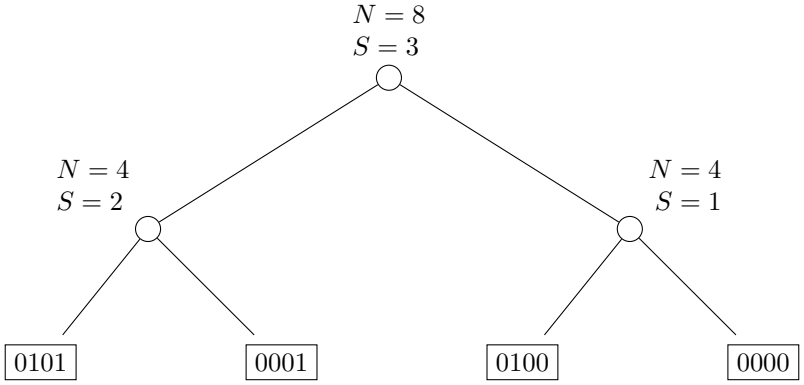


Figure 2.8: Dynamic bitset

Algorithm `access(`*node, i*`)`
  if `isLeaf(`*node*`)` **then**
    return node.bitset[i]
  end
  if node.N $\leq$ i **then**
    return `access`(node.left, i)
  end
  return `access`(node.right, i − node.N)

Algorithm 1: Accessing a value in a dynamic bitset

a	b	b	b	e	f	c	a	g	d	d
0	0	0	0	1	1	0	0	1	0	0
0	0	0	0	1	0	1	1 |	0	0	1
0	1	1	1	0 |	0	1	1 |	0	1 |	0

Table 2.2: Levelwise wavelet tree

Definition 11 (Select query on string)**.** *A select query $select_c(i)$ on a string S, computes the position of the ith occurrence of a character c in S.*

The classic data structure to efficiently compute *rank* and *select* queries on strings is the **wavelet tree** [18]. The data structure is a binary tree where every node consists of a bitset, and each level of the tree consists of the bits of a single position in each character of the string. Each bitset in the wavelet tree can for example be implemented as a dynamic bitset to allow insertions/deletions on a wavelet tree.

Figure 2.9 shows the wavelet tree for the string `abbbefcagdd`. The wavelet tree can perform *access*, *rank* and *select* queries by traversing the tree from root to leaf.

access(i) gives the bitstring of the character at a position i by first accessing position i in the root bitset, if the bit is a 0, we traverse to the left, if it is a 1 we traverse to the right. Before traversing to a child, we compute $rank_0(i)$ or $rank_1(i)$, depending on what bit is at position i, the resulting rank is the next position to consider in the subtree. This is done recursively

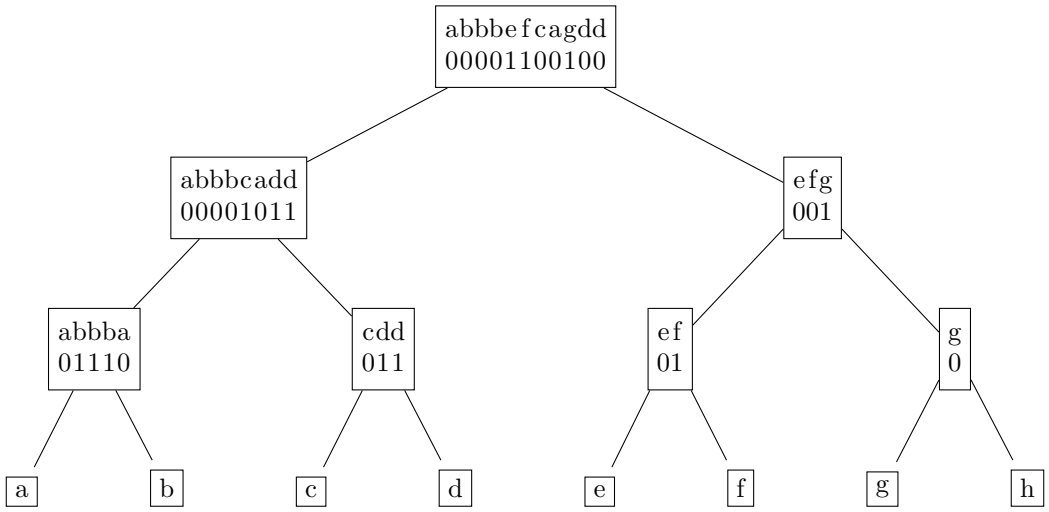


Figure 2.9: Wavelet tree for S = abbbefcagdd

until a leaf-node is found, and the bits which were examined at each node is the bitstring of the character at position i.

$rank_c(i)$ traverses the tree similarly to $access(i)$, but we use the bits of c to guide us to the correct subtree. When a leaf-node is reached, the value of i is returned, since when a leaf is reached, i will be pointing to the correct rank in the fictitious bitset of only c characters

The wavelet tree can be traversed in $O(\log \sigma)$ time where σ is the size of the alphabet. However, the *access*, *rank* and *select* operations also depend on how fast the bitsets can perform *access*, *rank* and *select* operations. If the bitsets are implemented using for example Jacobson's rank [24], the bitsets have a $O(1)$ *access*, *rank* and *select* time, but if we require dynamic bitsets for insertion/deletions, the time complexity increases to $O(\log n \log \sigma)$ because a bitset operation which takes $O(\log n)$ time is required in each level of the tree.

The wavelet tree can also be implemented without pointers, known as a levelwise or pointerless wavelet tree [31]. Table 2.2 shows the levelwise wavelet tree, which can be traversed level by level, but determining which interval of the bitset the node you are in occupies requires two calls to *rank*. The details of the levelwise wavelet trees are out of scope for this thesis, but leads us into the next wavelet data structure.

The **wavelet matrix** is a relatively recent improvement on the wavelet tree, which has been shown to be more efficient for *access*, *rank* and *select* queries, as well as less memory intensive for larger alphabets [12]. As the name suggests, the wavelet matrix is a matrix of bitsets, instead of a tree of bitsets. Similarly to the levelwise wavelet tree, for a string S of size n, each level of the matrix consists of a bitset of size n. The difference between a levelwise wavelet tree and a wavelet matrix is that the assumption that the "children" of an interval $[x, y]$ needs to occupy exactly $[x, y]$ is no longer true. We can instead put all 0 bits to the left in the next level, and all 1 bits to the right. For each level we store an integer z_l, which holds the number of zero bits in level l.

a	b	b	b	e	f	c	a	g	d	d
0	0	0	0	1	1	0	0	1	0	0
0	0	0	0	1	0	1	1	0	0	1
0	1	1	1	0	0	1	0	1	1	0

Table 2.3: Wavelet matrix for S = abbbefcagdd

Algorithm `access(`*WM, i*`)`

```
bits ← empty list
l ← 0
while l < Len(WM) do
    Add(bits, access(WM[l], i))
    if access(WM[l], i) = 1 then
        i ← z_l + rank_1(WM[l], i)
    else
        i ← rank_0(WM[l], i)
    end
    l ← l + 1
end
return bits
```

Algorithm 2: Wavelet matrix access

Algorithm $\texttt{rank}_c$`(`*WM, i*`)`

```
l ← 0
p ← 0
while l < Len(WM) do
    if access(WM[l], i) = 1 then
        i ← z_l + rank_1(WM[l], i)
        p ← z_l + rank_1(WM[l], p)
    else
        i ← rank_0(WM[l], i)
        p ← rank_0(WM[l], p)
    end
    l ← l + 1
end
return i − p
```

Algorithm 3: Wavelet matrix rank

For *access* operations, we traverse each level similarly to a levelwise wavelet tree, but instead of traversing a smaller and smaller interval in each level, we simply look at the current bit at position i, if it is a zero, the bit to examine in the next level is $rank_0(i)$, if the bit is a 1, the bit to examine in the next level is $z_l + rank_1(i)$. The $rank_c(i)$ operation is carried out similarly, but we also keep track of a preceding position to the final i, which we subtract from i, to count only the number of proceeding number of a characters. Table 2.3 shows an example wavelet matrix, and algorithm 2 and 3 shows how *access* and *rank* operations can be carried out for a wavelet matrix. The time complexities for *access*, *rank* and *select* operations is the same as for wavelet trees.

The wavelet matrix is constructed level by level. In level i, the bit at position i in each character is inserted. Before constructing the next level, each character is sorted by the current bit, so that all the characters with 0 bits at position i occupies the left side of level $i+1$, and vice versa for 1 bits.

Burrows-Wheeler transform

The Burrows-Wheeler transform (BWT) is a transform on strings, often performed to improve compression and searching [11]. The transform is computed by sorting all "cyclic-shifts" of the string lexicographically and extracting a new string from the last column of the cyclic-shift" matrix. The terminating symbol \$ is added to the string, which makes some algorithms on the BWT simpler. \$ is always the lexicographically smallest character. Table 2.4 shows the BWT for the string S = BANANA\$. There is a strong correlation between the BWT of a string and its suffix array, which is why this transform will be relevant for our detection algorithm.

Definition 12. *Cyclic-shift*

Index	CS
0	BANANA$
1	ANANA$B
2	NANA$BA
3	ANA$BAN
4	NA$BANA
5	A$BANAN
6	$BANANA

(a) Cyclic shifts

Index	CS
6	$BANAN**A**
5	A$BAN**AN**
3	ANA$BA**N**
1	ANANA$**B**
0	BANANA**$**
4	NA$BAN**A**
2	NANA$B**A**

(b) Sorted cyclic shifts and BWT

L	F
0	$Rank_A(0) + C[A] = 0 + 1 = 1$
1	$Rank_N(1) + C[N] = 0 + 5 = 5$
2	$Rank_N(2) + C[N] = 1 + 5 = 6$
3	$Rank_B(3) + C[B] = 0 + 4 = 4$
4	$Rank_\$(4) + C[\$] = 0 + 0 = 0$
5	$Rank_A(5) + C[A] = 1 + 1 = 2$
6	$Rank_A(6) + C[A] = 2 + 1 = 3$

(c) LF function

Table 2.4: S = BANANA$, BWT = ANNB$AA

Algorithm `ComputeBWT(`*S, SA*`)`

```
n ← S.len
BWT ← string of length n
for i from 0 to n do
    pos ← (SA[i] − 1) % n
    BWT[i] = S[pos]
end
return BWT
```

Algorithm 4: Computing the BWT of a string S from its suffix array

The cyclic-shift of order i for a string S is the cyclic-shift of S such that all characters in S are rotated i characters to the left. The cyclic-shift of order i is denoted CS_i

Definition 13. *Burrows-Wheeler transform*

The Burrows-Wheeler transform of a string S is a transformation on S where the cyclic-shifts of S are sorted, and the final character in each shift is concatenated.

As mentioned, there is a strong correlation between the suffix array of S and the BWT of S. Table 2.4 also shows that the indices of the sorted cyclic-shifts correspond to exactly the SA of S. This coincides because when sorting cyclic-shifts, everything that occurs after the $ of the CS will not affect its lexicographical ordering. This is because no CS will have a $ in the same position, so comparing two cyclic shifts lexicographically will always terminate whenever a $ is found. This means that the ordering of cyclic shifts is essentially the same as sorting all suffixes of S. Algorithm 4 shows how the BWT of S can be calculated directly from the SA of S in linear time. Since there is a one to one correlation between the SA and BWT, dynamic updates to a BWT would correspond to similar dynamic updates in the SA (and ISA). This will be useful in the upcoming implementation chapters for the incremental code clone detection algorithm.

An essential property of the BWT is that the transformation is reversible. By examining the BWT of a string S, we see that the BWT is a permutation of S. We will also see that there is a correlation between the characters in the first column of the cyclic-shift matrix and the last column, which allows us to compute the original string, given only the BWT.

S can be computed from only the BWT as long as there is a unique terminating character ($), or

the position of the final character is stored. Note that we only store the BWT here, which is the final column of the sorted cyclic-shifts. S is computed backwards by finding the final character (\$) in the BWT and then finding the CS where that character occurs in the first column (the previous CS). The character in the last column of that CS is $S[n-2]$. If there are multiple of the same character, the nth occurrence of a certain symbol in the last column will map to the nth occurrence of the same symbol in the first column. This process is repeated until we finish the cycle, returning to the final character (\$). Essentially, this process consists of traversing cyclic-shifts backwards and looking at the final character, which will give us S, since the final character of the current CS is continually shifting one position. We will use this technique in our detection algorithm not to compute the original input, but to be able to traverse our input string backwards.

Without storing the first column, it might seem hard to determine which CS to move to when traversing backwards. We can use the fact that the first column consists of the letters of S in sorted order to determine the previous CS with only the BWT. We can calculate the previous CS from only the last column by determining how many characters are lexicographically smaller than the current character, and also the rank of the current character at this position. The sum of these two values is the location of the previous CS. This function is called the Last-to-first function (LF function). Calculating the LF function can be done efficiently by using a rank/select data structure such as a wavelet tree [18] or wavelet matrix [12] to calculate the rank of all the characters in the BWT, and an array C which contains the number of occurrences of each letter in the BWT. In table 2.4, we can compute the original S from the BWT by starting at position 4, because that is the position of the \$, then use the LF-function, $LF(4) = 0$, which means the second to last character is $BWT[0] = A$. $LF(0) = 1$, which means the next character is $BWT[1] = N$, and so on.

The LF function will also be useful in the detection algorithm where we dynamically update the suffix array of S, given the changes of the BWT of S.

Chapter 3

Implementation: LSP server architecture

The following chapters will discuss how CCDetect-LSP is implemented, and especially focuses on our novel algorithm based on dynamic extended suffix arrays[1]. The current chapter will look at the LSP integration and how source code is initially indexed, while the following two chapters will first discuss how clones are initially detected and then how this analysis in incrementally updated as source code changes.

CCDetect-LSP is integrated into IDEs via LSP. The tool starts up as an LSP server, which an IDE client can connect to and send/receive messages from. The goal of the LSP client-server interaction is to give users of the tool an overview of all clones as they appear in source code, and allow them to navigate between matching clones. CCDetect-LSP is implemented in Java with the LSP4J library which provides an abstraction layer on top of the protocol which is easier to work with programmatically.

Figure 3.1 shows how clones are displayed in the VSCode IDE. The image shows a section of a file where a Java switch statement has been duplicated in two other files. Code clones are displayed inline in the file as an "Information" diagnostic, and when hovered over, the client shows the `DiagnosticRelatedInformation` information of the diagnostic where all the matching clones are displayed and can be clicked to navigate to the corresponding match. For a client which may not implement the `DiagnosticRelatedInformation` message[2], invoking a code-action to navigate to the matching clone is another option.

The following user-stories shows how interaction with the LSP server works.

- A programmer wants to see code clones for a file in their project, the programmer opens the file in their IDE and is shown diagnostics inline with the code wherever there are detected clones. The matching code clones are not necessarily in the same file.
- A programmer wants to see all code clones for the current project. The programmer opens

[1] CCDetect-LSP source code is available here: `https://github.com/jakobkhansen/CCDetect-lsp`

[2] Added to LSP in version 3.16, released in 2020. In our experience, some IDEs do not implement this feature.

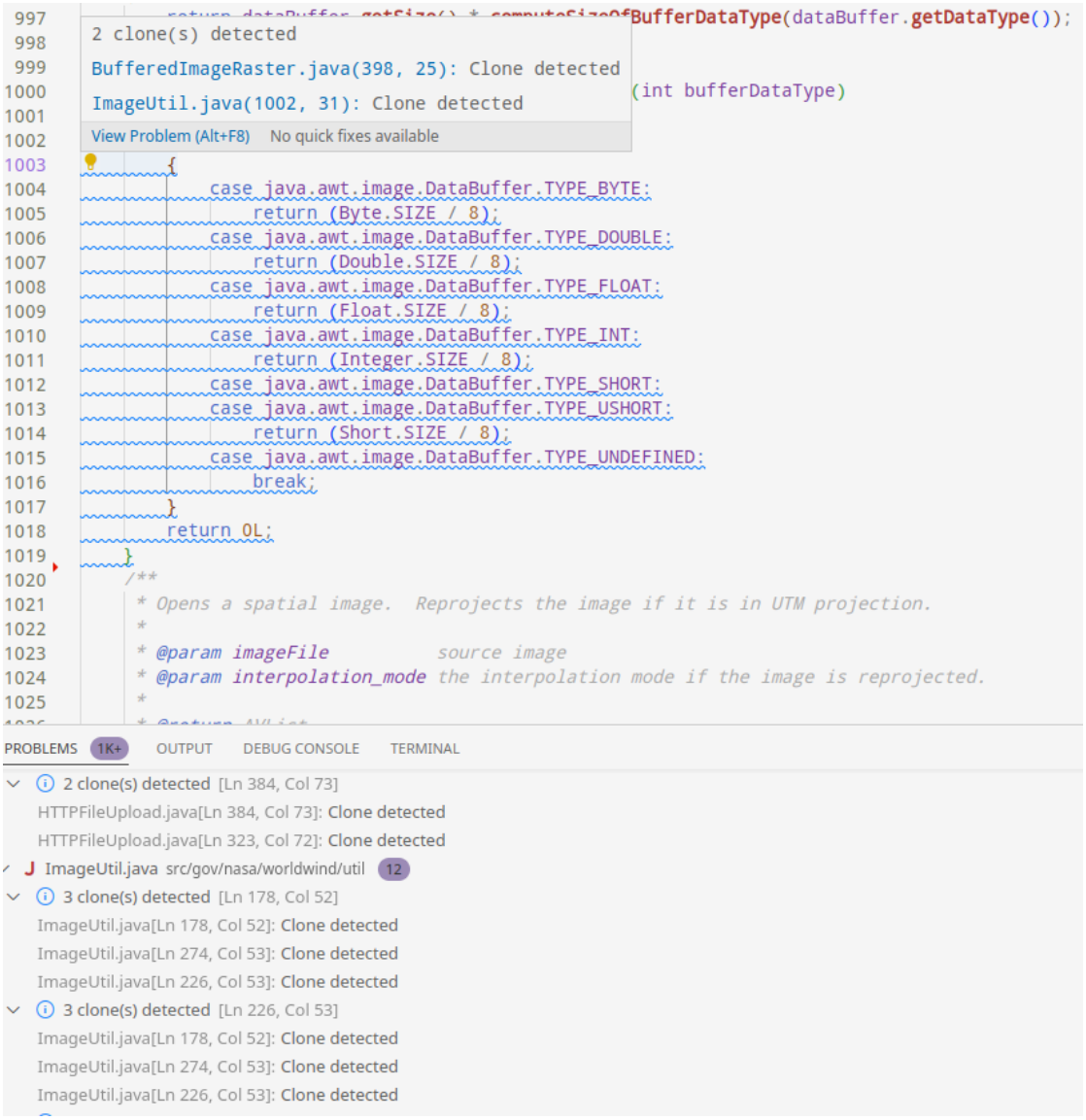


Figure 3.1: Code clones displayed in the VSCode IDE

the IDEs diagnostic view and will see all code clones detected as diagnostics there. The diagnostic will contain information like where the clone exists, and where the matching clone(s) are.

- A programmer wants to jump to one of the matches of a code clone in their IDE. The programmer moves their cursor to the diagnostic and will see a list of the matching code clones. The programmer selects any of the code clones which will open the file and location of the selected code clone. Alternatively, a code-action can be invoked to navigate, if the client does not implement the `DiagnosticRelatedInformation` interface.

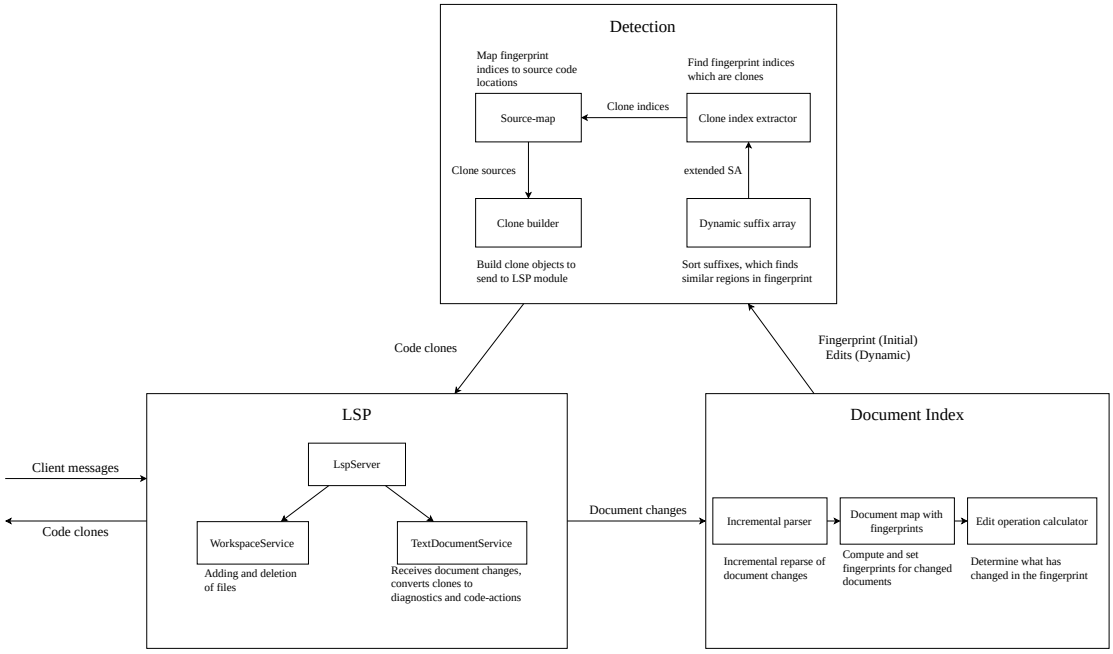


Figure 3.2: Tool architecture

- A programmer wants to refactor a set of clones by applying the "extract-method" refactoring (not implemented by CCDetect-LSP). The programmer performs the necessary refactorings, saves the file and will see that the refactored clones are no longer detected when the LSP server incrementally updates its analysis.

- A programmer is editing code and has written or copy-pasted a function which was already present in the source code of a large code base. As soon as the programmer saves the file, CCDetect-LSP will perform an incremental update based on the change, and display that the duplicated source code is a clone. The clone will be displayed until it is removed or modified.

- A programmer wants to switch to a project in a different programming language, and wants to use CCDetect-LSP to analyze clones in that project. The programmer edits the CCDetect-LSP configuration sent from the client to select the new language, which involves changing two parameters in the configuration. The configuration is explained in more detail in section 3.3. If the programmer wants to add a new language which has not yet been added in the CCDetect-LSP source code, a grammar for the language needs to be added before CCDetect-LSP can analyze that language.

- A programmer wants to use CCDetect-LSP in a new IDE. The process of running CCDetect-LSP in an IDE is different for each IDE. In general, an IDE which implements LSP, will have some sort of setup for launching an LSP server with a command, or on a certain event. The setup for CCDetect-LSP will be the same as for any other LSP server, but if extra configuration beyond the default values is wanted (such as changing language), the set of configuration options need to be sent from the client, as explained in section 3.3.

Figure 3.2 shows the architecture of the tool. The LSP module communicates with the IDE and delegates the work of handling documents to the document index. Detection of clones is delegated to the detection module. The LSP module receives a list of clones from the detection module which is then converted to LSP diagnostics and code-actions which is finally sent to the client.

3.1 Document index

Upon starting, the LSP server needs to index the project. This involves creating an index and inserting all the relevant documents in the code base. A document consists of the content of a file along with extra information such as the file's URI and some information which is useful for the clone detection. We define the following record for our documents:

```
Document {
    String uri,
    String content,
    AST ast,

    // Location in fingerprint
    int start,
    int end

    // Used for incremental updates
    int[] fingerprint,
    boolean open,
    boolean changed,
}
```

Each document in the index primarily consists of the contents, the URI and the AST of the document in its current state. Storing the AST will be useful for the incremental detection algorithm.

There are two things to consider when determining which files should be inserted into the index. First, we are considering only files of a specific file type, since the tool does not allow analysis of multiple programming languages at the same time (but multiple instances of CCDetect-LSP can be launched at the same time). Therefore, the index should contain for example only `*.java` files if Java is the language to analyze. Second, all files of that file type might not be relevant to consider in the analysis. This could for example be generated code, which generally contains a lot of duplication, but is not practical or necessary to consider as duplicate code, since this is not code which the programmer interact with directly. Therefore, the default behavior is to consider only files of the correct file type, which are checked into version control. The tool supports adding all files in a folder, or all files checked into version control.

When a document is first indexed by the server, the file contents is read from the disk. However, as soon as the programmer opens this file in their IDE, the source of truth for the files content is no longer on the disk, as the programmer is changing the file continuously before writing to the disk. The LSP protocol defines multiple JSON-RPC messages which the client sends to the server in order for the server to keep track of which files are opened, and the state of the content of opened files.

Upon opening a file, the client will send a `textDocument/didOpen` message to the server, which

contains the URI for the opened file. The index will at this point set the flag `open` for the relevant document and stops reading its contents from disk. Instead, updates to the file are obtained via the `textDocument/didChange` message. This message can provide either the entire content of the file each time the file changes, or it can provide only the changes and the location of the change. Receiving only the changes will be useful for this algorithm when the analysis incrementally reparses the document.

3.2 Displaying and interacting with clones

When detection is finished and the list of clones is ready, the LSP module will convert clones into diagnostics and code-actions which can be interacted with from the client. For diagnostics, each clone object is converted into a diagnostic which has a range and some information about the clone. The diagnostic displays an error in the client which also has information for each matching clone. This is achieved this by attaching `DiagnosticRelatedInformation` to a diagnostic which is defined in the protocol. When all the code clones have been converted to diagnostics, the server sends a `textDocument/publishDiagnostics` message which sends all the diagnostics to the client in JSON-RPC format. For code-actions, the process is similar, for each clone pair, we create a code-action which navigates from one to the other, using the `window/showDocument` message. These code-actions are similarly sent to the client via the `textDocument/codeAction` message, but this message is only invoked when the client specifically requests code actions at a certain location.

3.3 Configuration

CCDetect-LSP allows the LSP client to configure different parameters of CCDetect-LSP. The configuration works by utilizing the `intializationOptions` section of the `initialize` message, which is the first message sent from the client to the server. The client can send initialization options to the CCDetect-LSP server, with the following default values if none are given:

```
{
    language = "java",
    fragment_query = "(method_declaration) @method",
    clone_token_threshold = 100,
    dynamic_detection = true,
    update_on_save = true,
    ignore_nodes = [],
}
```

The fragment query option is a Tree-sitter query, which is explained in more detail in chapter 4.

Chapter 4

Implementation: Initial detection

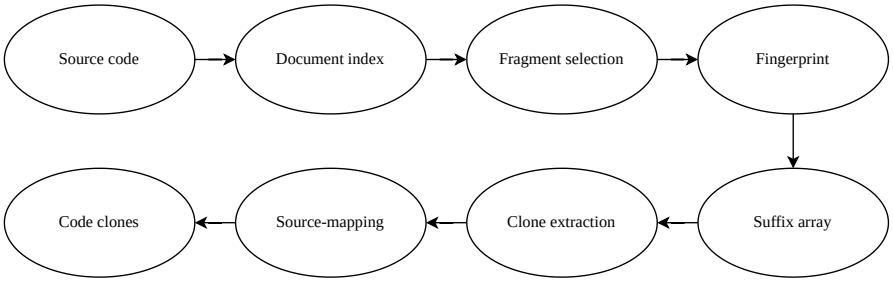


Figure 4.1: Phases of the detection algorithm

This chapter discusses the detection module of CCDetect-LSP and how the initial detection algorithm finds code clones. It consists of the detection algorithm which takes the document index as input, and outputs a list of code clones. The initial input to the algorithm will be the raw source code of each document in the index, in text format. Figure 4.1 shows all the phases of the detection algorithm and in each section a figure will be shown to illustrate which part of the pipeline is currently being discussed. The previous chapter has already detailed the first two phases, how a document index is built for each file of the source code.

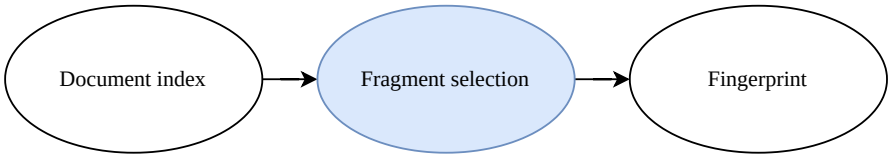
The algorithm of CCDetect-LSP detects syntactical type-1 code clones, based on a token threshold. Clones detected are therefore snippets of source-code which has at least N equal tokens, where N is a configurable parameter. The clones detected are exactly equal, but allows differences in for example white-space, and anything else which is not parsed into the syntax-tree. Additionally, Tree-sitter grammar often defines some AST nodes such as comments as "extra" nodes, which are ignored. There is also a configurable parameter to ignore additional nodes. There are also two rules which govern which clones are included in the list of reported clones.

- Clones which are completely contained within another larger clone are excluded.
- Clones cannot extend past the end of a fragment.

In broad strokes, the algorithm first selects the relevant parts of source code to detect code clones in (fragment selection), then transforms the selected fragments into a smaller representation (fingerprinting). For the matching, an extended suffix array for the fingerprint is constructed, where the LCP array is used to find long matching instances of source code. Finally, clones are filtered and aggregated into clone classes before they are source-mapped back to the original source-code locations, which the LSP server can send to the IDE in the form of diagnostics.

The decision to go with extended suffix arrays were based on the fact that extended suffix arrays use less memory than the traditional suffix tree approach, which is important in an IDE scenario. We also wanted to explore dynamic extended suffix arrays to determine if it could be a better approach to clone detection with regards to performance as well, as we are not aware of any related work which implements a dynamic extended suffix array and compares its performance with a dynamic suffix tree.

4.1 Fragment selection



The first phase of the algorithm involves extracting the relevant fragments of source code which should be considered for detection. A fragment, in this case is an instance of a particular type of node in the AST. The tokens the node encompasses (leaf-nodes of the subtree) is in this phase extracted and used in the subsequent phases. Since the algorithm is language agnostic, it is not feasible to have a single algorithm for fragment extraction or to define a separate fragment extraction algorithm for every possible language. Therefore, a parser generator tool is used, which can generate the code for parsing any programming language given a grammar. We use Tree-sitter, a parser generator with incremental parsing capabilities, as well as a query language for finding any type of node in the AST [10]. Tree-sitter queries are flexible queries to select specific nodes or subtrees. For example, in Java it would be natural to consider only methods. The Tree-sitter query for selecting only the method nodes in a Java programs AST would be:

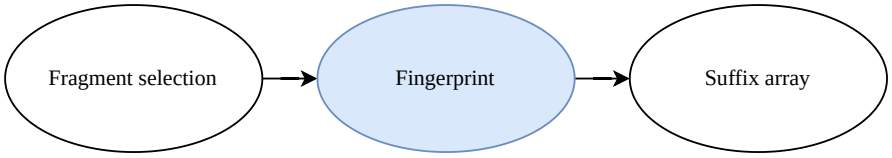
```
(method_declaration @method)
```
(4.1)

This Tree-sitter query selects the node with type `method_declaration` and "captures" it with the `@method` name, so that it can be further processed by the program. Readers interested in the details of the query system are referred to the Tree-sitter documentation [10].

Using a tree-sitter query, the algorithm parses the program into an AST and queries the tree for a list of all nodes which matches the query. For each node, all the tokens in the subtree of the node are extracted and sent to the next phase.

Implementing something similar for another parser/AST could be as simple as traversing the tree until a node of a specific type is found, using the visitor pattern [16, p. 366].

4.2 Fingerprinting




```
public class Math() {
    public int multiplyByTwo(int param) {
        return param * 2;
    }

    public int addTwo(int param) {
        return param + 2;
    }
}
```

(a) Source-code

Token	Fingerprint
public	2
int	3
multiplyByTwo	4
(	5
param	6
)	7
{	8
return	9
*	10
2	11
;	12
}	13
addTwo	14
+	15

(b) Fingerprint mapping

[2, 3, 4, 5, 3, 6, 7, 8, 9, 6, 10, 11, 1, 2, 3, 14, 5, 3, 6, 7, 8, 9, 6, 15, 11, 1, 0]

(c) Fingerprint, terminals underlined

Figure 4.2: Example fingerprint of Java source-code

The next phase of the algorithm is to transform the extracted tokens into a representation which is less computationally heavy for the matching algorithm. The goal is to reduce the total size of the input which needs to be processed.

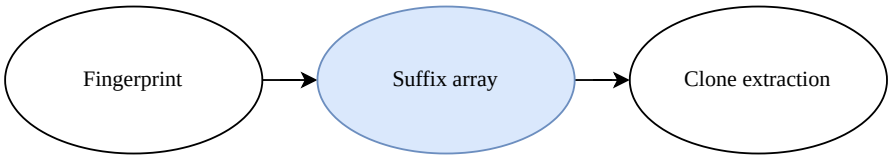
Because the algorithm that is used for matching is based on an extended suffix array, the representation should be in a format similar to a string, which is the standard input to a suffix array construction algorithm. However, it is not strictly necessary to use strings. The essential property of the input array is that we have a sizeable alphabet where each element is comparable. The required size of the alphabet is correlated with the number of unique token values in the code base. Therefore, it is safe to say that a larger code base requires a larger alphabet, as larger code bases will generally have more unique token values. We will use an array of integers instead of a string as it has a large alphabet in most programming languages. The size of the alphabet of `int` values in Java is large enough that it will not be the limiting factor for how large code bases can be analyzed.

The algorithm utilizes fingerprinting in order to reduce the size of the representation. Fingerprinting is a technique which involves taking some part of the input and mapping it to a smaller

bitstring, which uniquely identifies that part. In this case, the algorithm takes each token of the source-code, and maps its token value to the bitstring of an integer. The algorithm stores and increments a counter, which is used to fingerprint a token value whenever a new unique token is found. The mapping therefore starts with small integers, and uses larger and larger integers as more unique token values are encountered. The algorithm also stores a map from the token value, to the fingerprint value, which is used to map multiple occurrences of the same token value to the same fingerprint value.

Figure 4.2 shows how a sample Java program would be fingerprinted. Note that in the example only the tokens of the methods were extracted, which is why the class declaration is not fingerprinted. The fingerprint mapping initializes its counter at 2 to make space for some terminating characters. Each fragment is terminated by a 1 and the fingerprint is ultimately terminated by a 0. Having these values in the fingerprint will be useful for the matching algorithm where the suffix array is constructed and utilized for detection. The fingerprint of each fragment is stored in the relevant document object.

4.3 Suffix array construction



The next step is to input the fingerprint into a suffix array construction algorithm (SACA), so that the suffix array can be used to find repeating sequences. Recall that the suffix array of a string sorts all the suffixes of the string, as discussed in section 2.7. This makes it simple to find similar regions of text, since the most similar suffixes will be adjacent in the suffix array. All the fingerprints which are stored in each document object will now be concatenated to be stored in a single integer array (with terminators after each fragment), which the suffix array (SA), inverse suffix array (ISA) and longest-common prefix array (LCP) is computed from.

We have utilized a straight-forward implementation of the "Induced sorting variable-length LMS-substrings" algorithm [34] which computes a suffix array in linear time. The following section will give a high-level overview of how the algorithm works.

The algorithm will be assumed to have a string as its input, as this is most common for suffix arrays, and working with strings can be more clear to the reader when considering suffixes. The algorithm is still applicable to our fingerprint, since an array of integers will work similarly to a string when given as input.

The "Induced sorting variable-length LMS-substrings" algorithm (often abbreviated SA-IS) is an algorithm that works by divide and conquering the suffix array and inducing how to sort the suffixes of the string S from a smaller string, S_1. This smaller string consists of the "building blocks" of S and will make it simple to compute the rest of the suffix array of S. First, we will introduce some definitions and theorems used for the algorithm. Input to the algorithm is a string S with length n.

$$type(S,i) = \begin{cases} \text{S-type} & \text{if i is the last character of S} \\ \text{S-type} & \text{if S[i] < S[i + 1]} \\ \text{L-type} & \text{if S[i] > S[i + 1]} \\ type(S, i+1) & \text{if S[i] = S[i + 1]} \end{cases}$$

Figure 4.3: Suffix type

B	A	N	A	N	A	$
L	S	L	S	L	L	S
0	1	2	3	4	5	6

Table 4.1: Suffix types of S = BANANA$, LMS characters underlined

Definition 14 (L-type and S-type suffixes)**.** *A suffix starting at position i in a string S is considered to be L-type if it is lexicographically larger than the next suffix at position i+1, meaning* $\text{suffix}(S,i) > \text{suffix}(S,i+1)$. *Conversely, a suffix is considered to be S-type if* $\text{suffix}(S,i) < \text{suffix}(S,i+1)$. *The sentinel suffix ($) of S is always S-type.*

Note that two suffixes in the same string cannot be lexicographically equal, therefore all cases are handled by this definition. Determining the type of each suffix can be done in $O(n)$ time by scanning S from right-to-left and observing the following properties: $\text{suffix}(S,i)$ is L-type if $S[i] > S[i+1]$. Similarly, $\text{suffix}(S,i)$ is S-type if $S[i] < S[i+1]$. If $S[i] = S[i+1]$, then $\text{suffix}(S,i)$ has the same type as $\text{suffix}(S,i+1)$. This is true because if the first character of the current suffix is not equal to first character of the next suffix, we already know the type based on the first character. If the first character is equal, we have effectively transformed the problem to finding the type of the next suffix, since we are now comparing the second character of the current suffix, with the second character of the second suffix. Since we have already computed the type of the next suffix, we can reuse the value. Figure 4.3 shows a recurrence which determines the type of each suffix in a string in $O(n)$ time and table 4.1 shows an example.

Definition 15 (LMS character)**.** *An LMS (Left-most S-type) character in a string S is a position i in S such that $S[i]$ is S-type and $S[i-1]$ is L-type. An LMS-suffix is a suffix in S which begins with an LMS character. The final character of S (the sentinel) is always an LMS character and the first character is never an LMS character.*

Definition 16 (LMS-substring)**.** *An LMS-substring in a string S is a substring $S[i..j]$ in S such that $i \neq j$, $S[i]$ and $[j]$ are LMS characters and there are no other LMS characters between. The sentinel character is also an LMS-substring and is the only LMS-substring of length ≤ 3*

LMS-substrings form "basic-blocks" in the string S. Each LMS-substring is mostly lexicographically decreasing or increasing, which is easier to sort. Table 4.2 shows that in the string BANANA$ there are 3 S-type suffixes and all of them form LMS-substrings (ANA, ANA$, $). Using this notion, we can sort all suffixes recursively using the following theorems:

Theorem 2. *Given sorted LMS-suffixes of S, the rest of the suffix array can be induced in linear time.*

Buckets	$	A	B	N
Initial	{ -1 }	{ -1, -1, -1 }	{ -1 }	{ -1, -1 }
Slot LMS	{ 6 }	{ -1, 3, 1 }	{ -1 }	{ -1, -1 }
Slot L-type	{ 6 }	{ 5, 3, 1 }	{ 0 }	{ 4, 2 }
Slot S-type	{ 6 }	{ 5, 3, 1 }	{ 0 }	{ 4, 2 }

Mapping	$	0
	ANA$	1
	ANA	2

S_1	210
SA of S_1	[2, 1, 0]
Sorted LMS-substrings	[6, 3, 1]

Table 4.2: Building S_1 and sorting LMS-substrings of S = BANANA$

Theorem 3. *There are at most $n/2$ LMS-substrings in a string S of length n.*

These theorems are proven in the original paper [34], we will now use these theorems to demonstrate how we can compute the suffix array in linear time. Table 4.2 shows a running example of the algorithm. We can construct a smaller string S_1 by first sorting the LMS-substrings. Sorting LMS-substrings can be done by first bucket-sorting each LMS-substring by its first character. The buckets are implemented as arrays and each LMS-substring is inserted at the end of the correct bucket. Afterwards, L-type suffixes are "bucketed" by iterating over the buckets, and for each suffix in the bucket, insert the suffix to the left, if it is L-type. Meaning that if we encounter the suffix at position 6, the suffix at position 5 is bucketed if it is L-type. These suffixes should be inserted into the beginning of their respective bucket. Finally, S-type suffixes are bucketed in the same fashion, but the buckets are scanned from right-to-left and at positions where the previous suffix is S-type, the previous suffix is inserted at the end of its bucket. This final step could possibly change the ordering of the LMS-substrings.

After sorting, each equal LMS-substring is given a unique increasing integer value. Two LMS-substrings are considered equal if they are equal in terms of length, characters and types. S_1 is now built by mapping each LMS-substring to its unique value, and concatenating them in the original ordering. S_1 is now a smaller case which is used in the recursion. The string is at most $n/2$ in size, meaning there will be at most $\log_2(n)$ recursive calls. The recursive call will return the suffix array of the S_1. Nong et al. proves a theorem which shows that sorting S_1 is equivalent to sorting the LMS suffixes of S. Therefore, the suffix array of S_1 can be mapped to the LMS suffixes of S [34].

The base-case of the recursion is when the suffix array can be computed simply by bucketing each suffix, which happens when every suffix of the string starts with a unique character.

Table 4.2 shows how the LMS-substrings are bucketed and how S_1 is constructed. We see that the final buckets are actually equal to the final SA which we are trying to compute. This is because the LMS-substrings are already sorted in reverse order, which is not true for any arbitrary input. S_1 consists of only unique characters, therefore the suffix array of S_1 is computed by simply bucketing each suffix and returning the array.

```
Algorithm ComputeISA(SA)
    n ← Len(SA)
    ISA ← array of size n
    for i from 0 to n do
        ISA[SA[i]] ← i
    end
    return ISA
```

Algorithm 5: Compute ISA from SA

When the recursive call returns, the SA of S_1 is used to determine the order which the LMS-suffixes should be slotted into the bigger SA. By scanning the smaller SA and mapping those indices back to the indices of the original LMS-substrings, we get a sorted ordering of the LMS-suffixes. We then scan the sorted LMS-suffixes from right-to-left and slot each LMS-suffix at the end of its bucket, the rest of the L-type and S-type suffixes will be slotted correctly afterwards. For BANANA\$ this will be the exact same process as in table 4.2, since the LMS-substrings were already inserted in reverse ordering.

There will be $O(\log_2(n))$ recursive calls, where each recursive call takes $O(n)$ time, with n halving in each call. Therefore, the recurrence will have the complexity of:

$$T(n) = \begin{cases} O(n) & \text{if base-case.} \\ T(n/2) + O(n) & \text{otherwise.} \end{cases}$$

Which means $T(n) = 2n = O(n)$.

Building ISA and LCP arrays

Computing the ISA after constructing the SA is simple. Since the ISA is simply the inverse of SA, it can be constructed in linear time with a single loop as seen in Algorithm 5.

Computing the LCP in linear time is more complicated and requires some insight about which order to insert LCP values. We will also add one extra restriction to the LCP values, being that the LCP values cannot match past a 1, which were the terminal value between fragments. This restriction will be useful when we want to extract clones using the LCP array. The algorithm to compute the LCP is shown in Algorithm 6. The intuition for this algorithm is that if a suffix at position i has an LCP value l describing the common-prefix between it and some other suffix at position j, then the LCP value of the suffix at position $i + 1$ is at least $l - 1$, since there exists suffixes at position $i + 1$ and $j + 1$ which shares at minimum l characters, except for the first character which is cut off. Therefore, they share at least $l - 1$ characters, and the algorithm can start comparing the characters at that offset.

Now that we have computed the extended suffix array in linear time, we are ready to detect clones in our code base using this data structure.

Algorithm `ComputeLCP(`*S, SA, ISA*`)`

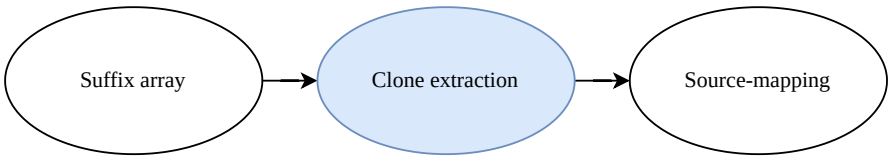
```
n ← Len(SA)
LCP ← array of size n
lcpLen ← 0
for i from 0 to n − 1 do
    r ← ISA[i]
    prevSuffix ← SA[r − 1]
    while S[i + lcpLen] = S[prevSuffix + lcpLen] and S[i + lcpLen] ≠ 1 do
        lcpLen ← lcpLen + 1
    end
    LCP[r] ← lcpLen
    lcpLen ← Max(0, lcpLen − 1)
end
return ISA
```

Algorithm 6: Compute LCP from input string S, SA, and ISA

Index	Suffix						Minimum LCP
1	A	N	A	N	A	$	0
3	A	N	A	$			0
2	N	A	N	A	$		2
4	N	A	$				2

Table 4.3: Minimum common LCP values between suffixes for S = BANANA$

4.4 Clone extraction



With the extended suffix array computed, we can now consider which substrings (prefixes of suffixes) we want to extract as potential code clones. In this phase the indices of the fingerprint which we consider to be code clones are extracted, which will be mapped back to the original source code in the next phase.

A straightforward solution is to extract every suffix which has an LCP value which is greater than the token threshold. The algorithm is a loop over S, using ISA to find the corresponding LCP value. This finds the clone indices, as shown in Algorithm 7.

However, this algorithm will return a lot of contained clones. A contained clone is a clone where all the tokens of the clone is a part of another, larger clone. The algorithm will give a lot of contained clones, for example in the case where there is a suffix with an LCP value of 100, the next suffix will have the LCP value of at least 99 and likely matches with the same code clone as the previous suffix, but with an offset of 1 token. For any large code clone, there will therefore be many smaller clones which are completely contained within it, but these clones are also likely to

```
Algorithm SimpleCloneExtraction(S, ISA, LCP)
    n ← S.len
    clones ← list
    for i from 0 to n − 1 do
        if ISA[i] = 0 then
            continue
        end

        if LCP[ISA[i]] ≥ THRESHOLD then
            Insert(clones, i) // Adds i to the clone-list
        end
    end
    return clones
```

Algorithm 7: Extract clones indices in a string S

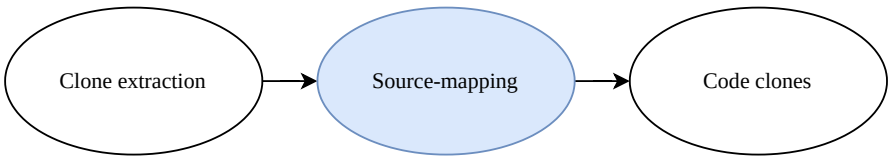
match with another clone which is also contained within the larger clones match. Since the code clones point at mostly the same area, the contained clones are not very useful, and is discarded by the tool. We extend our clone extraction algorithm to account for this, by using the following theorem:

Theorem 4. *The LCP of a suffix at position i is completely contained in the LCP of the previous suffix at position $i-1$ if the LCP value of the suffix at position $i-1$ is greater than the LCP value of the suffix at position i. Meaning* $\text{LCP[ISA[i]]} < \text{LCP[ISA[i} - 1]]$.

Algorithm 8 adds a while-loop to the clone extraction algorithm which skips over suffixes which are contained according to the theorem. Note that this algorithm doesn't disallow contained clones entirely, but any clone which is a shorter version of another clone pointing to the same match, will be discarded. Also note that overlapping clones, meaning two clones which share tokens, but where neither contains the other, will not be filtered.

At the end of this phase we now have a list of indices in the fingerprint which is considered to be code clones.

4.5 Source-mapping



With the clone indices in the fingerprint computed, we are almost finished. The final step is to map the clone indices back to the original source code. In order to correctly identify where a clone is located, we need to know which file the clone is located in, the range of the source code in that file the code clone covers, and the other matching code clones.

```
Algorithm CloneExtraction(S, ISA, LCP)
    n ← S.len
    clones ← empty list
    for i from 0 to n − 1 do
        if ISA[i] = 0 then
            continue
        end

        if LCP[ISA[i]] ≥ THRESHOLD then
            Insert(clones, i) // Adds i to the clone-list

            while i + 1 < n and LCP[ISA[i + 1]] < LCP[ISA[i]] do
                i ← i + 1
            end
        end
    end
    return clones
```

Algorithm 8: Extract clones indices in a string S, ignoring contained clones

To accomplish this, each document in the index needs to store the range of each of its tokens and keep track of which portion of the fingerprint consists of the documents tokens. This is done by storing two integer variables, each storing the start position and end position that the document has in the fingerprint.

To determine which document a fingerprint position corresponds to, we can perform a binary search on the list of the documents, which is sorted based on the start position of the documents fingerprint. The goal is to find the document D where the fingerprint position i is

$$D.start \leq i \leq D.end$$

Once the correct document has been found, we simply have to look up the correct range which the document stores. The index of this range is $i - D.start$.

Algorithm 9 returns the source-map of a given position i in the fingerprint, which includes the URI for the document and the source code range of the token at position i.

This algorithm only shows how to look up the position of a single token. Since a code clone is a range between two tokens, we have to look up the position at index i extracted in the previous phase, and the position where the clone ends, which is index $i + \text{LCP}[\text{ISA}[i]]$. The range of the code clone is therefore the combination of the starting range of the first token (at position i) and the ending range of the second token (at position $i + \text{LCP}[\text{ISA}[i]]$):

Aggregating clones

With this algorithm which gets the clone locations, the next step is to make sure that matching code clones are collected into buckets of clone classes. Remember that the LCP array only gives us the longest match between two suffixes, but it is naturally possible to have more than two

```
Algorithm SourceMap(documents, i)
    left ← 0
    right ← Len(documents) − 1

    while left ≤ right do
        mid ← (left + right)/2
        if documents[mid].end < i then
            left = mid + 1
        else if documents[mid].start > i then
            right = mid − 1
        else
            D ← documents[mid]
            range ← D.ranges[i − D.end]
            return (D.uri, range)
        end
    end
```

Algorithm 9: Get source-map for a position i in the fingerprint

clones of the same code snippet. This case happens when multiple consecutive indices in the SA are considered to be clones. Since we only look for type-1 clones, the transitivity property holds, meaning that if

$$SA[i] \xrightarrow{clone} SA[i+1] \xrightarrow{clone} SA[i+2]$$

then clones at position $SA[i]$ and $SA[i+2]$ are clones as well. This should be achieved by making sure that every new clone-pair discovered adds previously detected clones to their clone sets, and previously discovered clones also add new clones to theirs. Figure 4.4 shows how a new match is found in two previously disjoint clone classes, and the resulting aggregated clone class.

This is achieved in algorithm 10 where we build a clone-map, where the key is the index that a clone starts in the fingerprint, and the value is a code clone object. For each clone index i which was extracted in the last phase, we get the corresponding match index j (SA[ISA[i] − 1]), and for both i and j we look in the clone-map if there already is a clone at that position, or a new clone object is created and put in the map. Then, in order to aggregate the previously discovered clones and the new clone together, the set of matching clones is unioned between the two clones. In this way, all the previous existing clones of i is added as a clone in j and vice versa. Every clone in the two clone classes will then receive the same set of code clones. The `UnionCloneClass` function unions all the clone sets in both clone classes and then adds a match from all clones in one clone class to all clones in the other.

Finally, we have a list of every code clone in the code base, which is then sent to the LSP module of the tool, which handles the displaying of code clones to the client as shown in chapter 3.

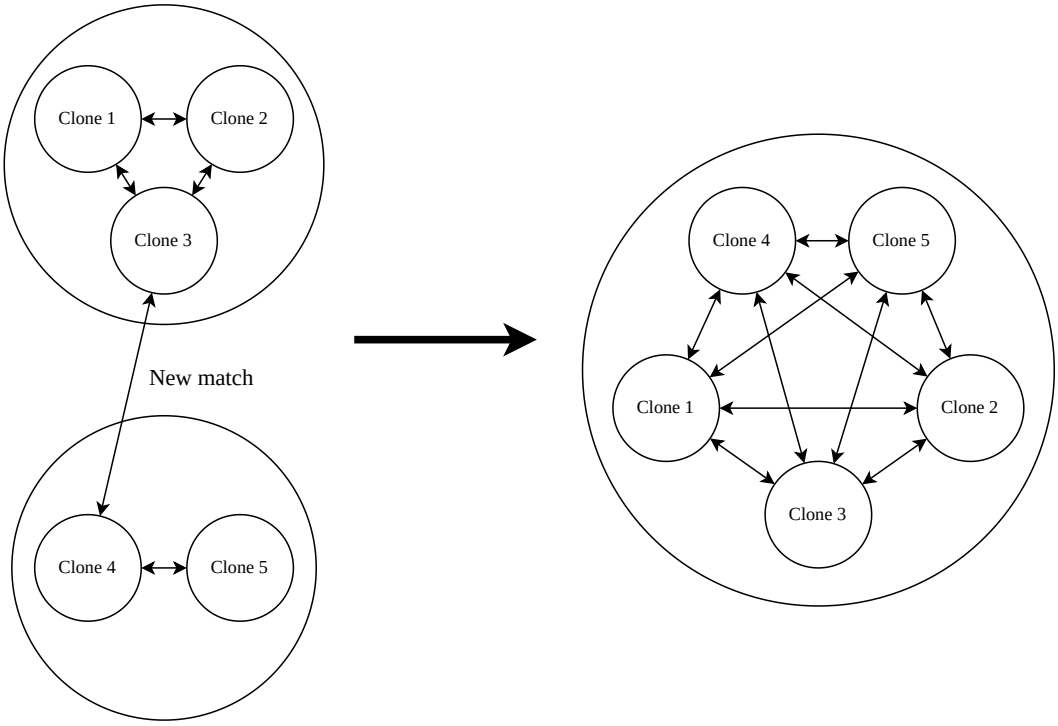


Figure 4.4: Clone class aggregation when a new match is found

```
Algorithm GetCloneMap(index, cloneIndices)
    n ← cloneIndices.len
    cloneMap ← Empty map with type: int → CodeClone

    for i from 0 to n − 1 do
        firstIndex ← cloneIndices[i]
        secondIndex ← SA[ISA[firstIndex] − 1]
        size ← LCP[ISA[firstIndex]] − 1

        // Use SourceMap function to get ranges and documents of clones
        firstRange ← Get source between firstIndex and (firstIndex + size)
        secondRange ← Get source between secondIndex and (secondIndex + size)

        firstDocument ← firstRange.document
        secondDocument ← secondRange.document

        // Build clone objects if they don't exist
        if firstIndex not in cloneMap then
            firstClone ← new CodeClone(firstDocument.uri, firstRange)
            Put(cloneMap, firstIndex, firstClone) // Put clone with firstIndex as key
        end
        if secondIndex not in cloneMap then
            secondClone ← new CodeClone(secondDocument.uri, secondRange)
            Put(cloneMap, i, secondClone) // Put clone with secondIndex as key
        end

        // Union clone classes
        UnionCloneClass(Get(cloneMap, firstIndex), Get(cloneMap, secondIndex))
    end
    return cloneMap
```

Algorithm 10: Build clone-map given the clone indices in the fingerprint

Chapter 5

Implementation: Incremental detection

The following chapter will present the incremental algorithm which efficiently updates the list of clones, without having to rebuild the full extended suffix array. Given an edit to a file in the project, we will be able to update the document index, fingerprint, extended suffix array and list of clones faster than the initial detection.

The incremental algorithm is run whenever a document is changed. The document index is signalized of a change either when a file is saved, or on any keystroke, configurable by the client.

In broad strokes, when a file is changed, we will first update the document index, which involves also incrementally parsing the changes to the file. From there we want to determine which edits have happened to the fingerprint, which we compute with an edit distance algorithm. The edits are the input to the dynamic extended suffix array algorithm, which is updated to reflect the extended suffix array of the new fingerprint. Afterwards, the clones are extracted from the extended suffix array similarly to the initial detection, but with slight optimizations based on storing positions of potential clone locations between revisions.

5.1 Affordable operations

Before discussing the algorithm, it is useful to determine the time cost associated with different operations and discuss what operations we can afford. The baseline we are comparing against is the initial detection where everything is computed from scratch, which runs in linear time in the size of the code base and fingerprint. We are therefore only looking to perform operations which are less computationally heavy than a linear scan over the code base. We can for example afford to iterate over the contents of a single file, which will be useful for our algorithm.

We cannot afford to iterate over the entire code base, meaning the contents of all files in the code base. The initial detection takes linear time in the size of the code base, therefore, iterating over the entire code base will likely make the algorithm as slow as the initial detection. The same goes for the iterating over the entire fingerprint.

As mentioned, we can afford to iterate over the contents of a single file. Iterating over the contents of a file with up to a few thousands lines is not a very expensive operation to perform and will not take a significant amount of time. The whole string of the file contents will be needed for Tree-sitter to incrementally parse the file.

Parsing an entire file however, can be too expensive in some cases. While the running time of parsing a single file is still linear in the size of file content, parsing a large file from scratch can take a significant amount of time in practice, and for small code bases with large files, parsing can take a significant portion of the total running time of an incremental update. This is why incremental parsing with Tree-sitter is used, which lowers the runtime closer to $O(|\text{edit}|)$, rather than $O(|\text{file}|)$

We can also afford two more useful operations, iterating over the documents in the index (not their contents), and iterating over the clones. We can afford to do these operations because the number of documents and the number of clones is likely multiple magnitudes smaller than the contents of the entire code base. Iterating over the documents will be useful when we are updating the document index, and iterating over the clones is necessary to do when the clones are mapped back to their original source code locations.

Why we can afford these operations will be become clearer in chapter 6, the main idea is that all the operations we can afford will take an insignificant amount of time compared to updating the extended suffix array.

5.2 Updating the document index

The first step of the incremental algorithm is to update the document index. We will also look at how we can reduce the memory usage of the index without a loss in terms of time complexity.

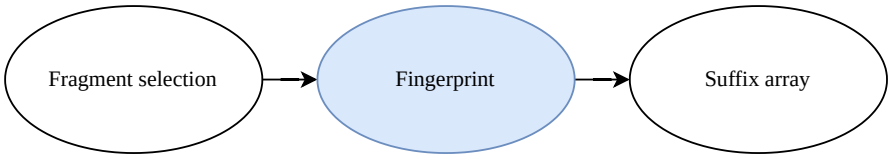
Each document stores its own content, AST and fingerprint. It is not strictly necessary to store either the content or the AST in memory all the time, as it is likely that only a handful of files are open in the IDE at once. Therefore, in the initial detection, we can free the memory of the file content and AST for each document after the fingerprint has been computed. However, if a file is opened in the IDE, the file can now be changed, so we should facilitate efficient updates for these files at least. When a file is opened, the file content should be read from the disk and updated via the `textDocument/didChange` messages sent from the client. It is also important to keep the AST of the opened file in memory in order to facilitate incremental parsing of the opened files.

When a file is opened, the LSP client sends a `textDocument/didOpen` message to the server, which finds the relevant document D in the index, and sets the following fields:

$$
\begin{aligned}
D.\text{open} &= \textbf{true} \\
D.\text{content} &= \texttt{Read}(D.\text{uri}) \\
D.\text{AST} &= \texttt{Parse}(D.\text{content})
\end{aligned}
$$

After the document fields have been set, the document is ready to receive updates. When the LSP client sends a `textDocument/didChange` message, the message consists of the URI of the edited file, the range of the content which has changed, and the content which has potentially been inserted. This range is then used in a Tree-sitter incremental parse of the file content. After this parse, we have efficiently updated a documents content and AST. After this update, we also set $D.\text{changed} = \textbf{true}$.

5.3 Updating fingerprints



With the updated AST for all documents, we can update the fingerprint of all documents which have been changed. For each document D where $D.\text{changed} = \textbf{true}$, the fingerprint for D may have changed. Computing the new fingerprint is the same process as in the initial detection, where we first query the AST for all nodes of a certain type, then for each matched node N, we extract and fingerprint all the tokens which N covers, using the same fingerprint mapping as was used for the initial detection.

An additional change we have to consider when incrementally updating fingerprints is that for a document D, $D.start$ and $D.end$ which corresponds to the range which D covers in the fingerprint, may have changed. Also, any document D_1 where $D_1.start > D.start$ could also have its range changed. This range needs to be updated in order for the source-mapping to work. This is solved while updating each documents fingerprint by counting the number of tokens in each document after updating, and setting the appropriate `start` and `end` fields.

Figure 5.1 shows how an index of three documents is updated when two tokens are inserted. Note that each document stores its own fingerprint, we do not need to concatenate them to one large array as in the initial detection.

5.4 Computing edit operations

With the updated fingerprint, we could build the suffix array from scratch and already see a substantial improvement in performance, as is shown in chapter 6. The major bottleneck of the initial detection is to parse and fingerprint the entire code base. In order to improve the algorithm even further, we will dynamically update the extended suffix array to avoid recomputing it fully.

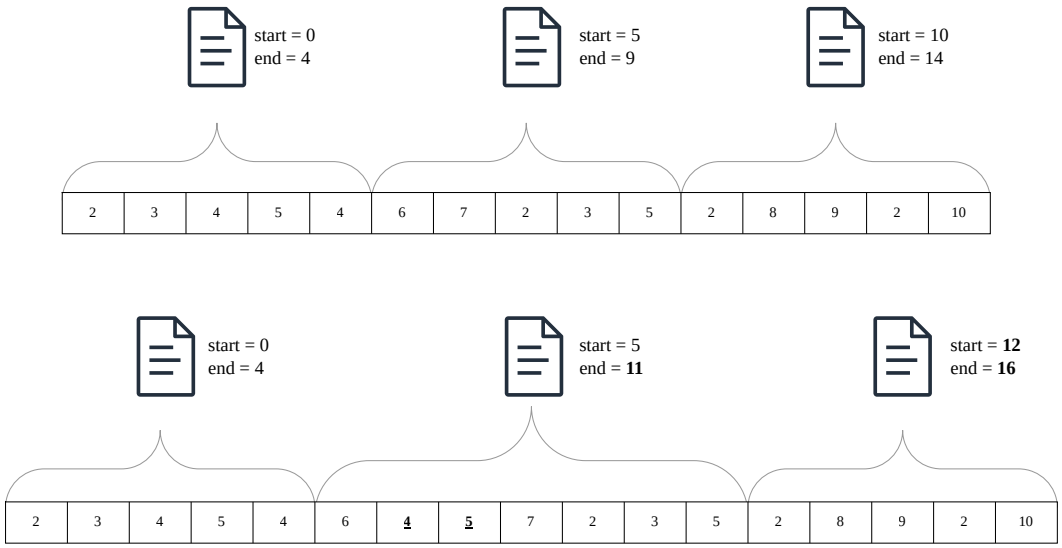


Figure 5.1: Old (above) and new (below) fingerprints and ranges when two tokens are inserted.

The input to the dynamic extended suffix array algorithm is a set of edit operations. An edit operation can either be deleting, inserting or substituting a set of consecutive characters in the fingerprint. The dynamic extended suffix array algorithm will take each edit operation, and update its state to reflect the new fingerprint. Therefore, the first problem to solve is to determine what exactly has changed in the fingerprint.

There are a couple of approaches we could take to determine the edit operations. The simplest is that whenever a file has changed, regard the entire file as changed, and perform a delete operation which removes the entire fingerprint of that file, and then an insertion which inserts the new fingerprint of that file. This is a simple approach which can have decent performance if the file is small, but can also lead to unnecessarily large edit operations. The fingerprint of the file is likely to be very similar to the previous version, therefore deleting and inserting the entire fingerprint could be a lot more work than it needs to be. However, this idea that we can always reduce the number of operations to only two, a deletion and an insertion, will be useful in our final algorithm.

Another possible approach is to look at the ranges that the LSP client provides with each `textDocument/didChange` message and determine which tokens in the fingerprint have been affected according to this range. However, this approach tightly couples the algorithm to LSP and the scenario where we know the exact ranges of each change. Also, we might do unnecessary amounts of operations if we do multiple edits, since some operations could cancel each other out, for example by inserting and then deleting the same text.

A better approach is to determine the changes of the fingerprint using an edit distance algorithm. An edit distance algorithm is an algorithm which computes the distance between two strings S_1 and S_2. Distance between two strings is the minimum number of edit operations (insert, delete, substitute) which is required to transform S_1 into S_2. Many of the algorithms which computes the edit distance, also allows computing what the operations are.

$$\sum_{i=0}^{n} M[i][0] = i$$

$$\sum_{j=0}^{m} M[0][j] = j$$

$$M[i][j] = \begin{cases} M[i-1][j-1] & \text{if } S_1[i] = S_2[j] \\ 1 + \texttt{Min}(M[i-1][j-1], M[i][j-1], M[i-1][j]) \end{cases}$$

Figure 5.2: Edit distance recurrence

		D	E	M	O	C	R	A	T
	0	1	2	3	4	5	6	7	8
R	1	1	2	3	4	5	5	6	7
E	2	2	1	2	3	4	5	6	7
P	3	3	2	2	3	4	5	6	7
U	4	4	3	3	3	4	5	6	7
B	5	5	4	4	4	4	5	6	7
L	6	6	5	5	5	5	5	6	7
I	7	7	6	6	6	6	6	6	7
C	8	8	7	7	7	6	7	7	7
A	9	9	8	8	8	7	7	7	8
N	10	10	9	9	9	8	8	8	8

Table 5.1: Edit distance matrix for REPUBLICAN → DEMOCRAT

The classic algorithm for calculating edit distance operations is attributed to Wagner and Fischer [45]. The input to the algorithm is two strings S_1 and S_2 of length n and m. The output will be the set of operations needed to turn S_1 into S_2. This algorithm is based on dynamic programming where a matrix M is filled from top to bottom and then the operations are inferred from M. The recurrence in equation 5.2 shows how the edit distance matrix is filled and table 5.1 shows an example matrix.

Each index i, j in M contains the edit distance value between the substrings $S_1[0..i]$ and $S_2[0..j]$ The values in M are calculated by determining what is the cheapest operation to do at a certain location to make the substrings equal. This can be determined by looking at the three surrounding indices in M: $M[i-1][j-1]$, $M[i-1][j]$ and $M[i][j-1]$. Each of these indices equate to deleting, inserting or substituting a character in S_1.

The edit operations can then be inferred from M by backtracking from the bottom-right index, to the top-left, giving us the edit operations in reverse. At each position i, j we choose either of the 3 surrounding indices, the same indices which were used to determine the value originally. Choosing the left index $(i, j-1)$ equates to inserting the character $S_2[j-1]$ at position $i-1$. Choosing the top index $(i-1, j)$ equates to deleting the character $S_1[i-1]$. Choosing the top-left index $(i-1, j-1)$ equates to substituting $S_1[i-1]$ with $S_2[j-1]$. If these characters are already

equal, the operation can be ignored. For example in table 5.1, the first operation is to substitute `R` with `D` at position 0. Afterwards `P` and `U` is deleted at position 2. Then the `B` which is now at position 2 is substituted by `M`. This continues with more substitutions until we finally have `DEMOCRAT`.

Aggregating edit operations

In the next phase we will feed the edit operations into an algorithm which dynamically updates our suffix array based on those operations. However, this algorithm will be more efficient if the operations are combined to singular inserts, deletes or substitutes of strings more than one character. The edit distance algorithm outputs only single character operations, meaning we insert, delete or substitute a single character at a time. We therefore want a way to combine these operations to "larger" operations.

We define an EditOperation with the following record:

```
EditOperation {
    OperationType type,
    char[] chars,
    int position
}
```

where type is either an insert, delete or substitute.

One way to combine operations is to find operations of the same type which are sequenced in the matrix. For example in table 5.1, we have two consecutive delete operations, where `P` and `U` is deleted at position 2. These two operations could be combined to a single delete operation of two characters.

The idea for the algorithm which computes the edit operations is to traverse the optimal matrix path backwards and for each operation we either append to the current operation if possible, or start a new operation. We can continue the current operation if the next operation is of the same type, and the next operation has the same position as the current operation. For example in table 5.1, we have two delete operations in a sequence at position 1 where `P` and `U` is deleted. The first operation we encounter is the deletion of `U`. At this point we create a new `EditOperation` with position 1 and `U` in the `chars` array. The next operation is also a delete operation at position 1, so we add the `P` to the list of characters for the operation to delete. The next operation is a substitute at position 0, so we cannot continue the delete operation, and a new substitute operation is created instead. Similarly, at position 2 to 5, we substitute `BLIC` with `MOCR`. This operation is computed in a backwards fashion similarly to how we did the deletion, except that the position of the operation is decremented for each character we add to it.

Note that this algorithm is not an optimal algorithm, as this is a trade-off between processing many characters in few operations, versus processing many operations with few characters. As mentioned earlier, one could always reduce the solution to be only two operations, deleting the whole string, and inserting the new string. This is only two operations, but likely processes more characters in total compared to our algorithm. See Appendix 18 for detailed pseudocode of the algorithm explained in this section.

		F	I	N	I	S	H	I	N	G
	0	1	2	3	4	5	6	7	8	9
F	1	0	1	2	3	4	5	6	7	8
A	2	1	1	2	3	4	5	6	7	8
S	3	2	2	2	3	3	4	5	6	7
C	4	3	3	3	3	4	4	5	6	7
I	5	4	3	4	3	4	5	4	5	6
N	6	5	4	3	4	4	5	5	4	5
A	7	6	5	4	4	5	5	6	5	5
T	8	7	6	5	5	5	6	6	6	6
I	9	8	7	6	5	6	6	6	7	7
N	10	9	8	7	6	6	7	7	6	7
G	11	10	9	8	7	7	7	8	7	6

Table 5.2: Edit distance matrix for FASCINATING → FINISHING. Blue is optimal path, green is minimized matrix

Optimizing memory usage

A problem with the above solution is the memory usage of the matrix. It is not feasible to input the entire old and new fingerprint into an edit distance algorithm, as the full fingerprint can have millions of symbols, and the old and new fingerprint is likely approximately the same size. This would require a matrix which is too large to fit in memory. We will use a few techniques to reduce the memory usage of this algorithm without compromising on the time complexity.

The first technique is to not input the whole fingerprint. We can drastically reduce the size of the input by only comparing the old and new fingerprint of the document D which has been edited. D stores its previous and current fingerprint, and whenever it is edited, we can compute the edit operations of D, with its previous and current fingerprint as input. This also avoids having to construct the full fingerprint by concatenating each documents fingerprint, which we had to do in the initial detection. With this approach however, the position of each edit operation of D will not have the correct position relative to the entire fingerprint, so D.start is added to the position of each edit operation to correct this.

Another optimization we can do to reduce the size of the matrix is to remove the "trivial" part at each end of our matrix. If we compare two strings which have many similar characters at the beginning or end of our string, we know that these will not produce any edit operations in the matrix, as they will simply be diagonal moves (substitutes) where the characters are already the same. In table 5.2 we see the edit matrix for the input `FASCINATING` and `FINISHING`. The two words share the common prefix `F` and the common suffix `ING`. If we examine the edit distance values, we see that the highlighted green matrix starts at 0 in the top left, and ends with 6 in the top right, which is the same final values as in the full matrix. Also, we see that there are no edit operations being added outside the green matrix. Using this knowledge, we can see that we only need to compare the strings `ASCINAT` and `INISH`, which would give us the exact same edit distance. We can also get the exact same edit operations, as long as we account for the starting offset of the original string, so that the operations have the correct positions. Algorithm 11 shows how two input strings can be minimized for usage in the edit distance algorithm. After getting the edit operations from the algorithm, the `startOffset` is added to the position of each operation to account for the offset of the minimized matrix.

```
Algorithm EditDistanceMinimizeStrings(S_1, S_2)
    for i from 0 to Min(Len(S_1), Len(S_1)) do
        if S_1[i] ≠ S_2[i] then break
    end
    startOffset ← i

    s1End ← Len(S_1) − 1
    s2End ← Len(S_2) − 1
    while s1End ≥ startOffset and s2End ≥ startOffset and S_1[s1End] = S_2[s2End] do
        s1End ← s1End − 1
        s2End ← s2End − 1
    end

    miniS1 ← S_1[startOffset...s1End]
    miniS2 ← S_2[startOffset...s2End]

    return (miniS1, miniS2, startOffset)
```

Algorithm 11: Minimize strings for edit distance algorithm

The two optimizations drastically reduce the memory and time usage of the edit distance algorithm, but in cases where a very large file is edited, and the beginning and end of the old and new fingerprint do not match, we can still encounter instances of the matrix being too large to fit in memory. The problem is that the matrix size has a polynomial growth in terms of the fingerprint size. This is because the old fingerprint is of size n and the new fingerprint is of size m, where $n \approx m$, which requires an $n \times m$ size matrix to calculate the edit operations. For example if a Java file contains 3000 lines, the number of tokens can exceed 10000, which would require approximately a 10000^2 size matrix, which is approaching a memory usage which is too much for an IDE scenario.

A solution to this problem is to reduce the required memory from the polinomial $O(n \times m)$ memory usage, to a linear $O(n)$ memory usage. A stepping stone towards such a solution is an observation attributed to Ukkonen, which reduces the required memory to linear growth in the size of either of the strings [43]. The observation shows that in order to compute the next row/column of the edit distance matrix, we only need the previous row/column. For example in table 5.3, the fifth row has been computed using only the fourth row. This is done by first computing the left-most index of the fifth row, which is always one more than the previous row. Afterwards, we can compute the other elements of the row from left to right, with the same recurrence, shown in equation 5.2. This is possible because we always know the above, left, and top-left elements of the current index. This can be implemented as two arrays, which holds the previous and current row.

This change allows us to compute the edit distance in linear space, but now the problem is how to find the actual edit operation. This is not possible, because we don't have the entire matrix available to traverse anymore. However, this can be solved using Hirschberg's algorithm [21]. Hirschberg's algorithm is an algorithm which can compute the edit operations of two strings in the same time complexity as the Wagner-Fischer algorithm, but uses only linear space in the size of either of the input strings.

		D	E	M	O	C	R	A	T
R									
E									
P	3	3	2	2	3	4	5	6	7
U	4	4	3	3	3	4	5	6	7
B									
L									
I									
C									
A									
N									

Table 5.3: Edit distance matrix with minimal memory usage

Hirschberg's algorithm

The first insight we need for this algorithm is that there is at minimum one edit operation on each row/column of the edit distance matrix. This is intuitive, because in order to "travel" from the top-left to the bottom-right of the matrix, the path needs to visit at least one cell from the top to the bottom, visiting all the rows, and from the left to the right, visiting all the columns. Hirschberg's algorithm uses this insight to compute one position of the optimal path at a time, which it finds in the middle row between two already known positions of the path. Table 5.4 shows an example of how the positions are found.

Initially, we know that the top-left index, and the bottom-right index of the matrix are guaranteed to be part of the optimal path, since those positions are the starting and ending point of the path. The next position to find is in the middle row of the matrix. Determining which position in the middle row should be selected is done by performing the edit distance recurrence twice, once from the beginning to the middle row in the same fashion as the original algorithm, and once in reverse, from the bottom-right to the middle row. Both of the recurrences can run in linear space, because we only need to store two rows at a time. Summing the two results for the middle row gives us an array which can intuitively be understood as how long the minimal path which goes through the corresponding position in the row is. Since we know that there is at least one of the positions in this row which has to be part of the optimal path, the minimum value in this array corresponds to one position which has to be part of the optimal path. Once we know the middle position which is part of the optimal path, we can recursively call the function twice, once with the top-left and middle as input, and once with the middle and bottom-right as input. The base-case of the recursion is when the size of the matrix is linear in the size of either of the strings, in which case we call the Wagner-Fischer edit distance recurrence with a linear space complexity. Note that there can be multiple minimum values in the row where a position is selected, which corresponds to the fact that there can be multiple optimal paths through the matrix.

Each position which is a part of the optimal path is stored as they are found, and can then be traversed backwards to determine the edit operations similarly to traversing the matrix. For example if we iterate over the list of positions backwards, we will know that if the first position is at x, y and the second position is at $x - 1, y$, that corresponds to a delete operation, just as in the matrix-based algorithm.

		F	I	N	I	S	H	I	N	G
F										
A										
S										
C										
I	10	8	6	7	6	7	8	8	10	12
N										
A										
T										
I										
N										
G										

		F	I	N	I	S	H	I	N	G
F										
A	5	3	3	4	6					
S										
C										
I										
N										
A										
T					5	4	3	4	6	8
I										
N										
G										

		F	I	N	I	S	H	I	N	G
	0	1	2							
F	1	0	1							
A	2	1	1							
S			2	2	4					
C										
I										
N					3	3	4			
A										
T										
I							2	0	2	4
N										
G										

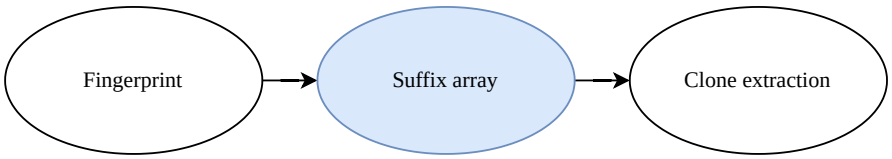
		F	I	N	I	S	H	I	N	G
F										
A										
S				0	1					
C				1	1					
I				2	1					
N					0	1	2			
A					1	1	2			
T					2	2	2			
I								0	1	2
N								1	0	1
G								2	1	0

Table 5.4: Hirschberg's algorithm. Blue cells are part of the optimal path, dark-blue cells are new positions.

Using Hirschberg's algorithm, we are now able to compute the edit operations for very large files without memory usage issues.

A final problem we should be aware of is that in some instances, the number of edit operations can explode. This happens in cases where there are no long stretches of similar operations in the matrix. As mentioned, our algorithm which aggregates edit operations together is not optimal, and could likely be improved to reduce the number of edit operations when we allow more than one character in each operation. Therefore, in instances where the edit distance algorithm returns so many edit operations that it is likely very slow to use them in the next phase, we will instead reduce the solution to only two operations, deletion and insertion of the entire minimized string. In practice this optimization can often avoid large spikes in the running time of the next phase.

5.5 Suffix array incremental updates



With the edit operations which describes how the fingerprint has changed, the next phase is to input the edit operations into an algorithm which dynamically updates the extended suffix array. The algorithm discussed in this section will be run once for each edit operation.

Constructing/updating the extended suffix array is the most time-consuming phase of the algorithm, so the goal for the dynamic update is to take the previous suffix array and an edit operation, and compute the new suffix array faster than building the suffix array from scratch. We will also in a following section convert the suffix array into a dynamic structure, which allows insertions, deletions and increments multiple elements faster than linear time.

The algorithm used is the algorithm presented by Salson et al. [38, 39]. This algorithm is presented as two algorithms, one which dynamically updates the BWT and one which dynamically updates an extended suffix array based on the BWT changes. We will first discuss how the BWT is updated and how that relates to updates in the suffix array. In the next section, we will present a new data structure which is more efficient for representing the suffix array in a dynamic setting, and then explain how the LCP array can be attached to this data structure. The original paper also describes how to incrementally update the ISA, but this will be omitted, as our new data structure contains the ISA implicitly.

Recall that the BWT is a transform on an input string which is often used for compression and text-search purposes. The BWT is also reversible by using the LF function to compute the original string in reverse. The LF function is computed for a position i such that $LF(i) = rank_{BWT[i]}(i) + C[BWT[i]]$ where the array C contains for any character c, how many characters are lexicographically smaller than c. This function can intuitively be used in the context of the BWT to allow us to move from CS_i to CS_{i-1}. Also recall that the suffix array and the BWT of a string is highly correlated to the point where one can compute the BWT from the suffix array in a single pass. Any updates to the BWT will also correspond to similar updates to the suffix array, so the algorithm which updates the BWT can be used to determine how the suffix array

Order	Cyclic-shift
6	$BANAN**A**
5	A$BANA**N**
3	ANA$BA**N**
1	ANANA$**B**
0	BANANA**$**
4	NA$BAN**A**
2	NANA$B**A**

(a) S = BANANA$

Order	Cyclic-shift
7	$BABNAN**A**
6	A$BABNA**N**
1	ABNANA$**B**
4	ANA$BAB**N**
0	BABNANA**$**
2	BNANA$B**A**
5	NA$BABN**A**
3	NANA$BA**B**

(b) S = BABNANA$

Table 5.5: BWT for string before and after insert

changes as well.

Updating BWT on insertion

We will observe which parts of a BWT changes when we insert a single character. The approach for deletions and substitutions are similar, and this approach can easily be extended to edit operations of more than one character. Note that we are only storing the BWT in a data structure, not the whole cyclic-shifts. We will consider an insertion of a character c, inserted at position i. When c is inserted, we know that we will have exactly one more CS for the string, which starts with the inserted character c, at position i. This CS ends with the character at position $i-1$. We also know that exactly one CS will change its last character, which is the CS previously starting at position i, but is at position $i+1$ after the insertion. This new CS will end with the inserted character, c. These are the only changes that happen to the characters in the BWT, but as we will see, the ordering of the BWT characters may not be correct anymore.

For the string `BANANA$`, let $c =$ `B`, and $i = 2$, which results in the string `BABNANA$`. In table 5.5 we can see that some substrings of the BWT is preserved, such as `$AA` and `AN`, but some other changes have occured. In table 5.6 We see that the new CS_2 is `BNANA$BA`, and the changed CS_3 is `NANA$BAB`. We can find where these cyclic-shifts are located by first looking up i in the ISA, ISA[2] = 6. This is the location of CS_3 (after the insertion), where the final character is updated to the inserted character `B`. To find the position where the new cyclic-shift should be inserted, we can use the LF function to move from the CS_3 to CS_2. Computing $LF(6)$ after the substitute gives us 5, which is where the new CS_2 should be inserted in the BWT. When we insert in the BWT, we only need to consider the final character of CS_2, which in this case is `A`, the character which was previously substituted.

The final stage of the algorithm now that all the elements are present, is to rearrange the cyclic-shifts which have changed their lexicographical ordering. It is possible that the cyclic-shifts change their ordering because of the new character. For example `ANANA$B` < `ANA$BAN`, but `ABNANA$B` > `ANA$BABN`. A useful observation is that only CS_j where $j \leq i$ will have their lexicographical ordering changed. This is because only in these cyclic-shifts the inserted character c comes before the `$`. Since the `$` is the smallest lexicographical character, and no two cyclic-shifts will have `$` in the same location, we know that the lexicographical comparison of two cyclic-shifts will never go past the `$`.

Order	F	L
6	\$	A
5	A	N
3	A	N
1	A	B
0	B	\$
4	N	A
2	N	A

(a) Original BWT

Order	F	L	
7	\$	A	
6	A	N	
4	A	N	
1	A	B	
0	B	\$	
2	**B**	**A**	Inserted
5	B	A	
3	**N**	**B**	$A \rightarrow B$

(b) After change and insert

Order	F	L	
7	\$	A	
6	A	N	
1	**A**	**B**	↕
4	**A**	**N**	
0	B	\$	
2	B	A	
5	B	A	
3	N	B	

(c) After reordering

Table 5.6: BWT for string before and after insert

We know that only CS_j where $j \leq i$ can have their ordering changed, and we have already inserted CS_i at the correct position (the new cyclic-shift). We will store two positions, *pos* and *expected*, where *pos* is initially the position of CS_{i-1}, which was stored before any changes were made to the BWT. *pos* is therefore the actual position of CS_{i-1}, *expected* is the expected position of CS_{i-1}, which is computed by computing $LF(insertionPoint)$ where *insertionPoint* is the position where the new CS_i was inserted. With the knowledge of where the cyclic-shift of order $i-1$ is, and where we expect it to be, we can move the cyclic-shift to that position. In the example, we have that $expected = 3$, and $pos = LF(5) = 2$ after the insertion gives us the expected location of the cyclic-shift of order 1. Therefore, we remove the cyclic-shift at position 3, and insert it again at position 2.

Before we move the character, we also compute $newPos = LF(pos)$ which will give us the position of the cyclic-shift of order $i-2$. After the move, we can continue comparing the position and expected position by updating $pos = newPos$, and $expected = LF(expected)$. In the example this would update both *pos* and *expected* to 4. Since the expected position and actual position is the same, the algorithm is done, and we know that every position of the BWT is now in the correct location. We know that there will be no more characters which are out of place for any other CS down to order 0 because $LF(pos) = LF(expected)$, which would be true for all future iterations as soon as $pos = expected$

Updating SA on insertion

The suffix array is updated similarly to the BWT. When we insert the new CS at position 5 in the BWT, we similarly insert the value 2 at position 5 in SA. We insert the value 2 because the new suffix which corresponds to the new cyclic-shift starts at the position where the new character was inserted, which was 2. When the new CS is inserted, we are now in an invalid state for the suffix array, because we have two elements with the value 2. Note that a suffix array is always a permutation of the values $(0..n)$ where n is the number of characters in the input. Since we have inserted a new suffix in the suffix array which is the 2nd smallest suffix, all the previous suffixes which had a value $j \geq 2$ is now incremented. After incrementing all these suffixes, we are now in a valid suffix array state, but similarly to our BWT, some suffixes may have had its ordering changed. For every reordering operation in the BWT, the elements in SA are reordered in the same fashion.

Algorithm 12 dynamically updates a suffix array and BWT when a single character is inserted.

Algorithm `UpdateSuffixArrayInsert`(*SA, ISA, BWT, i, ch*)

```
posFirstModified ← ISA[i]
previousCS ← LF(BWT, i)

storedLetter ← BWT[posFirstModified]
BWT[posFirstModified] ← ch

insertionPoint ← LF(BWT, i)
if storedLetter < ch then insertionPoint ← insertionPoint + 1

// Insert storedLetter in BWT at pointOfInsertion
Insert(BWT, insertionPoint, storedLetter)

// Insert i in SA at pointOfInsertion, increment all values ≥ position
Insert(SA, insertionPoint, pos)
IncrementGreaterThan(SA, pos)

if insertionPoint ≤ previousCS then previousCS ← previousCS + 1

pos ← previousCS
expected ← LF(insertionPoint)
while pos ≠ expected do
    newPos ← LF(pos)

    // Delete value at pos and reinsert at expected in BWT and SA
    MoveRow(pos, expected)

    pos ← newPos
    expected ← LF(expected)
end
```

Algorithm 12: Update BWT and suffix array when inserting a single character

This algorithm can easily be extended to insert a string instead of a single character by adding a loop which continuously updates the `pointOfInsertion` to the previous CS with the LF function, and inserts all the characters at that position into the BWT and SA backwards. This is more efficient than calling the single character algorithm multiple times, because the reordering stage is only performed once. The details for insertions/deletions of multiple characters is covered in the original paper [39].

Deletion

A deletion of a single character is a similar procedure as an insertion, as it is simply reversing the substitution and insertion, and then performing the reordering stage again. If we have the string `BABNANA$`, and delete the `B` at position 2, We know that there will be exactly one CS removed. This is CS_2 which starts with the deleted character `B`, and ends with the character before it, `A`. There is also a single CS which will have its final character changed, CS_3, because the final character `B` will be deleted. The final step is to again perform the reordering stage for

CS_j where $j \leq 2$, as these are the only cyclic-shifts which can have their lexicographical ordering changed.

The algorithm which performs the deletion at position i will do the deletion and substitution in the opposite order. First we find both the CS which will have its character substituted in the BWT, and the CS which will have its character deleted in the BWT. The character to be substituted will be found with $substituted = ISA[i+1]$, and the character to delete will be found with $deleted = LF(substituted)$. $BWT[deleted]$ is then deleted, and $BWT[substituted]$ is substituted with the character that was deleted. Finally, the reordering phase is performed in the same fashion, where *pos* is set to the position of the original CS_{i-1} and *expected* is set to $LF(substituted)$

In our example $substituted = ISA[3] = 7$, $deleted = LF(7) = 5$ and $pos = LF(5) = 2$. Then $BWT[deleted]$ is deleted, and $BWT[substituted]$ is set to the deleted character. Then $expected = LF(6) = 3$. Note that *substituted* was decremented after the deletion because the deletion moved that CS one position up. When the reordering stage happens, we have $pos = 2$ and $expected = 3$, so we perform the reordering in the BWT and SA in the same fashion as previously, and the algorithm will again terminate after one iteration of the reordering phase.

Substitution

Substitution operations are also possible and covered in the original paper, but in CCDetect-LSP these were simply implemented as a deletion followed by an insertion. This is less efficient than a single operation, since the reordering stage is done twice, but we do not suspect that this will worsen the performance in any significant way.

LF function

The LF function is used multiple times for any suffix array update. Therefore, it is important to be able to efficient compute LF at any point in time. The naive approach of linearly iterating through the BWT to compute the *rank* and number of smaller characters is too slow. Recall that $LF(i) = rank_{BWT[i]}(i) + C[BWT[i]]$ where $rank_{BWT[i]}(i)$ is the rank for the character $BWT[i]$ at position i in the BWT, and $C[BWT[i]]$ counts the number of lexicographically smaller characters than $BWT[i]$ in the BWT. We therefore need a data structure for *rank* queries on the BWT, and a data structure which stores the number of smaller characters of each character in the BWT.

For the *rank* queries, recall that the wavelet matrix was a data structure which could efficiently compute *rank*/*select* queries for strings. The data structure is implemented as a matrix of dynamic bitsets, where we also stored the number of 0 bits in each level. Traversing this matrix from top to bottom and performing rank queries on the dynamic bitsets in each level allows us to perform access, rank and select queries efficiently for a string input. We will use the wavelet matrix to store the BWT and perform efficient *access* and *rank* queries on it as needed. With the wavelet matrix, we do not need to store the BWT itself, as we can simply query the wavelet matrix to access characters of the BWT. We also do not need to store the full fingerprint, since we can access characters in the wavelet matrix and if we need to access the characters in a sorted order, we can use the LF function. We decided to use the wavelet matrix instead of a wavelet tree as it has been shown to have faster *rank* queries, and is more memory efficient for larger alphabets. Our alphabet can be quite large compared the standard applications of *rank*/*select* data structures (such as DNA analysis). The wavelet matrix is also generally simpler to implement, and can be dynamically updated in a simple manner. Each bitset in the wavelet

```
Algorithm WaveletMatrixInsert(wm, index, value)
    numBits ← Number of bits in value

    while numBits > Len(wm.levels) do
        Insert new row at level 0 in wm
    end

    level ← 0
    currIndex ← index

    while level < Len(wm.levels) do
        currentBit ← Len(wm.levels[level]) − level − 1

        // Get the bit value at position currentBit in value
        bit ← GetBit(value, currentBit)

        // Insert the bit into the bitset at position currIndex
        Insert(wm.level, currIndex, bit)

        // Update currIndex to the position to insert in the next level
        currIndex ← rank_bit(wm.level, currIndex)

        // If the bit we inserted is 1, add the number of zeroes in the level
        if bit = 1 then
            currIndex ← currIndex + wm.level.z_l
        end
        level ← level + 1
    end
```

Algorithm 13: Dynamically insert an element into the wavelet matrix

matrix is a dynamic bitset where we can efficiently insert and delete characters as needed when the size of the BWT grows/shrinks.

When inserting a character c into the wavelet matrix at position i, we insert the first bit of c in level 0 in the matrix at position i. The next bit of c is then inserted in level 1 in the matrix at position $rank_x(WM[0], i)$ where x is the value of the previous bit. We continue this process, updating the position by performing a rank query, and then inserting the bit at that position, until we reach the bottom of the matrix. Similarly, for a deletion, we find each bit which was inserted for that position in each level, and remove them. Algorithm 13 shows how an element can be dynamically inserted into the wavelet matrix. Algorithms for deletions and substitutions are left out, but follow the same idea. Substitutions are simply implemented as a deletion followed by an insertion.

A complication with the wavelet data structures is to update the data structure as the alphabet increases in size. In our case the alphabet size will at some point increase to a size where elements need an additional bit. For the wavelet matrix, this can be solved by simply inserting a new row at the beginning of the matrix (level 0), where the bitset consists of all zeroes. Afterwards, the new element is inserted in the normal fashion.

The C array can be implemented in two ways. Either $C[c]$ holds the number of smaller character than c in the input, or $C[c]$ holds the number of occurrences of c. There are trade-offs with either implementation, as storing the number of smaller characters requires a linear scan whenever a character is inserted or deleted, while storing the number of occurrences requires a linear scan to compute the number of smaller characters whenever the LF function is called. We have chosen the approach of storing the number of occurrences, as updating the array is simpler in cases where the alphabet size increases. When we later update LCP values, we will see that there is a possibility that choosing the option of storing the number of smaller characters in C is faster.

5.6 Dynamic extended suffix arrays

A major bottleneck when inserting or deleting elements in the suffix array is that when an element is for example removed, all elements after it needs to be moved by one index so that the gap is closed. This is also the case in the reordering stage, where all the elements between the old position and new position of an element needs to be moved by one position. In addition, since a suffix array is always a permutation from 0 to $n-1$ where n is the number of elements in the input, we need increment/decrement all elements greater than or equal to the element which was inserted or deleted. For example, if we insert a new element into the suffix array which indicates that a new suffix is the ith smallest, then the previous suffix with value i in SA should now have the value $i+1$, which applies to all other suffixes lexicographically greater than the new suffix as well. In a standard representation for a suffix array (an array), this would require either a linear scan through the entire SA, or a linear scan through ISA from position i, which is also linear in the worst case. Multiple linear scans through the SA would make the algorithm slower than the linear time SACA when the amount of inserted or deleted character increases.

In this section we will introduce a data structure we call the dynamic extended suffix array. This data structure consists of a dynamic permutation, which facilitates insertion and deletion of elements in a permutation from 0 to $n-1$, without linear scans to close gaps or incrementing/decrementing elements. The data structure stores SA, ISA and LCP and is inspired by the data structure introduced for the same purpose by Salson et al. [39]. The data structure used by Salson et al. uses the same dynamic permutation tree, but reduces the amount of values which is stored to reduce memory consumption, but this compression also slows down the access time of elements in the permutation. Our data structure will not be compressed, which allows faster access to arbitrary elements, and will be extended to also include the LCP values as well. Therefore, this data structure encompasses the entire extended suffix array.

The data structure consists of two balanced binary trees, with pointers between elements in one tree to the other. Figure 5.3 shows the two trees which are labeled the A tree and the B tree. We will see that these trees can intuitively be understood as a tree containing the indexes (A), and a tree containing the values (B) of our permutation. Note that the trees do not actually hold the values displayed in the figure, the values in each node of the figure only represents the inorder position of each node, which is called the inorder rank of the node.

To construct a permutation of length n such as $[6, 5, 3, 1, 0, 4, 2]$, we can insert a node in each tree for each element. After the trees both contain n nodes, pointers are added from a node in A to a node in B and vice versa (doubly linked). If an element in SA has a value i at position j, pointers will be added between the node in A with inorder rank i, and the node in B with inorder rank j.

Now the tree is complete, but we can also insert new nodes into the data structure as the suffix array grows. When a new value j is added at position i, a node is inserted into A such that it has the inorder rank of i, and a node is inserted into B such that it has the inorder rank of j. For example, if we wanted to insert a new value 4 at position 1 in the permutation in figure 5.3, we would insert a new node in A so that it is the right child of the node labeled 0, and then insert a new node in B which is the left child of the node labeled 4. The trees can intuitively be understood such that the inorder values in A represent the indices of the permutation, and the inorder values in B represent the values in the permutation. Therefore, if we look at a node in A with inorder rank i, and follow the pointer to a node in B with inorder rank j, this means that the permutation has the value j at position i.

Deleting a node at position i is similar, but we will instead find the node with a rank of i in A, then delete that node and the node it points to in B. As the trees are implemented as balanced trees, traversing, inserting and deleting nodes takes $O(\log n)$ time.

An essential property of this tree is that when a new value j is inserted into the permutation at position i, all indices $\geq i$ and values $\geq j$ is incremented by 1. This is because all the nodes which had an inorder rank greater than or equal to the new nodes in each respective tree has now been incremented by 1 simply because the structure of the tree has changed. No additional work is required to increment the elements.

The part we are still missing is how to find a node with inorder value i, and how to determine the inorder value of a given node. We cannot afford to traverse all the nodes in an inorder fashion to find a node with a specific inorder rank. Therefore, the trees are implemented as order-statistic trees [13, p. 340]. An order-statistic tree is a tree where each node x stores an additional integer $size$, which contains the number of nodes in the subtree rooted at x. This allows us to easily determine how many nodes is in the left subtree of a given node, and use this to determine where the node with a certain inorder rank is located. If we wanted to find the node in the A tree of figure 5.3 with inorder rank 4, we would start at the root, see that the left subtree of the root contains 3 nodes. Therefore, the roots inorder rank is 3. With this information we know that we need to traverse to the right child, and look for the node with inorder rank $4 - 4 = 0$ in that subtree (since we have already seen 4 nodes). Since the node labelled 5 has a rank of 1 (in its own subtree), we know to traverse left, and at that point we have reached the correct node.

Algorithm 14 shows how to find a node in a tree by its inorder rank, and algorithm 15 shows how to determine the rank of a given node. The algorithm to find the value at a certain position is simple with these algorithms. To find the value in the permutation at position i, find the node with inorder rank i in A, follow its pointer to a node in B, and then find the inorder rank of that node, which is the value at position i in the permutation. We can also find the position of any value j in the permutation by performing this algorithm in reverse, by instead starting at the B tree, finding a node with rank j, then following its pointer to a node in A and returning that nodes inorder rank.

With the ability to both find a value at a given position in the permutation, and also finding the position of a given value, we can use this data structure to represent the suffix array and inverse suffix array, instead of a simple array. Inserting, deleting and accessing a node in this data structure takes $O(\log n)$ time, and we no longer need a linear scan through the data structure to increment/decrement values, as values are automatically incremented/decremented when a new element is inserted or deleted.

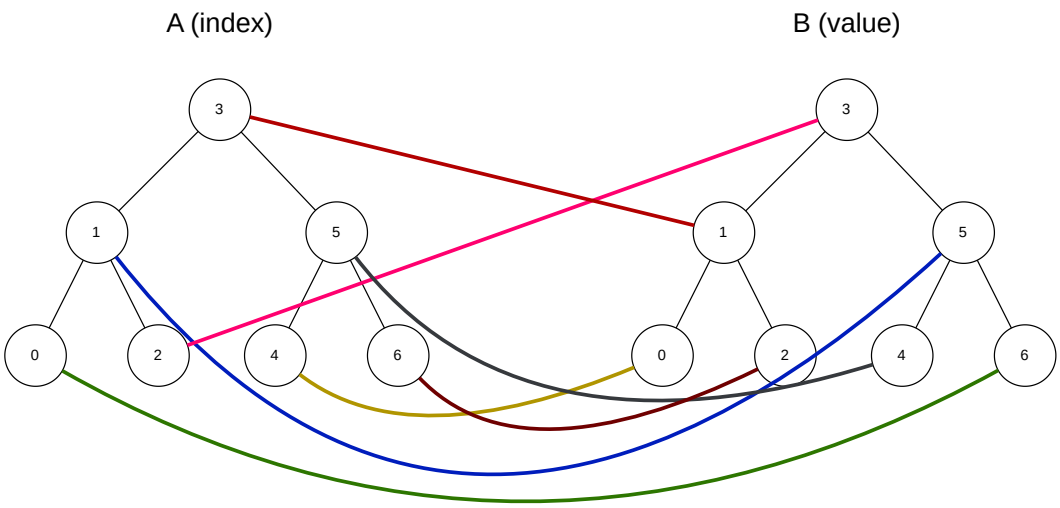


Figure 5.3: Dynamic permutation for the permutation $[6, 5, 3, 1, 0, 4, 2]$. Colors have no meaning except to make lines clearer

5.7 Dynamic LCP array updates

Now that we have a dynamic structure to represent SA and ISA, we can extend the data structure to also contain the LCP array, and efficiently update the LCP values as well.

Since every node in the A tree of our data structure correlates to an index, we can easily extend the data structure to contain LCP values by setting a value in each node of the A tree. This would make the A tree work as a balanced binary tree for the LCP values, and if an SA value at position i is deleted, the LCP value stored in the respective node is also deleted.

In the initial construction of the dynamic permutation, LCP values would be inserted together with the nodes. When dynamically updating the LCP values as suffixes are inserted/removed, the procedure is more complicated. Salson et al.[39] also describe a procedure for updating the LCP values after an insert/delete operation, which we will use.

When inserting a character into the input string, the suffix array changes by inserting a new value, and afterwards moving values from one position to another (reordering stage). Similarly, a deletion in the suffix array consists of deleting a value, and performing the same reordering operations. The moving of a value from position i to j can be decomposed into a deletion at position i followed by an insertion at position j. When updating the LCP array we therefore need to consider how insertion of a value and deletion of a value in SA affects the LCP values.

The idea for this algorithm is to keep track of all positions which could have its LCP value changed, and update the LCP values after the SA is fully updated. There are four different cases where an LCP value at a position i could be changed:

1. A suffix was inserted at position $i - 1$ in SA, which is the new suffix that the suffix at position i should compute its LCP value for.

2. The suffix at position $i - 1$ was deleted in SA, making the suffix at position $i - 2$ the new

```
Algorithm GetNodeByRank(root, rank)
    current ← root
    while current ≠ null do
        currentRank ← current.left.size
        if currentRank = rank then
            break
        else if currentRank > rank then
            current ← current.left
        else
            current ← current.right
            rank ← rank − currentRank + 1
        end
    end
    return current
```

Algorithm 14: Find the node in a tree with a given inorder rank

```
Algorithm GetRankOfNode(node)
    rank ← node.left.size
    while node.parent ≠ null do
        if node.parent.right = node then
            rank ← rank + node.parent.right + 1
        end
        node ← node.parent
    end
    return rank
```

Algorithm 15: Find the inorder rank of a given node

suffix that the suffix at position i should compute its LCP value for.

3. A character was inserted or deleted in the middle of the LCP for the suffix at position i in SA, which shortens the LCP value between it and the suffix at position $i-1$, and the LCP value between the suffix at position i and $i+1$.

4. A character was inserted or deleted at the end of the LCP for the suffix at position i in SA, which could potentially extend the LCP value between it and the suffix at position $i-1$, and the LCP value between the suffix at position i and $i+1$.

To cover the first two cases, we will store a dynamic bitset where a set bit at position i indicates that at some point, the LCP value at position i needs to be updated.

For an insertion at position $i-1$ in SA (case 1), we will set the bit at position i, as the suffix it compares its LCP to has changed. We will also insert a set bit at position $i-1$, as the newly inserted suffix also needs to compute its LCP value with the suffix at position $i-2$.

Similarly, for a deletion at position $i-1$ in SA (case 2), we will set the bit at position i, as the suffix it used to compare its LCP to has been deleted. We will also delete the bit at position $i-1$, as the suffix at that position has been deleted.

Algorithm `DynamicPermutationAccess(`*permutation, i*`)`
aNode = `GetNodeByRank(`permutation.A.root, i`)`
bNode = aNode.pointer
return `GetRankOfNode(`bNode`)`

Algorithm 16: Get value at position i in a permutation

After setting these bits while performing an insert/delete operation in the SA, the bitset now contains all the positions of LCP values which need to be updated for the first two cases. The other two cases are more complex to find the positions of. Instead of inserting the positions of these suffix into the bitset, we will iterate over the suffixes which could possibly have their LCP value changed. Iteration over every suffix and recomputing the LCP value takes linear time in the size of the fingerprint and is therefore too slow for this algorithm. However, we can reduce the number of suffixes we need to update by applying a series of observations.

The first observation is that we only need to consider suffixes where the insertion/deletion of a character at position i has changed the suffix. Those are the suffixes which start at a position j where $j \leq i$. Other suffixes cannot change, as the insertion/deletion happened before the starting position of the suffix. For a suffix at position j which has changed, two LCP values can change, the values at position k and $k+1$ where $SA[k] = j$. These two LCP values can change, because at position k, the suffix at position $SA[k]$ is compared with $SA[k-1]$, and at position $k+1$, the suffix at position $SA[k+1]$ is compared with $SA[k]$. Since both of these LCP values depend on the suffix at position $SA[k] = j$, they can potentially change when the suffix is changed. The positions in SA which possibly can be changed can be found using the LF function. In the final phase of the suffix array update algorithm when we have inserted or deleted an element at position i, we are reordering suffixes at position $j \leq i$. These are the suffixes we are looking for, but we can skip all suffixes which are reordered, as they are already marked to be computed in the bitset for the previous cases. Therefore, the variable *pos* points to the next suffix we need to consider at the end of algorithm 12.

Another useful observation when iterating over suffixes is that after an insertion/deletion, no two suffixes will have the insertion/deletion occur at the same position. This is true because no two suffixes start on the same position in the input, and an insertion/deletion at a position i in the input can therefore never correspond to an insert/deletion at the same position for two different suffixes. With that in mind, another useful observation is that given two suffixes with an LCP value of p, any change inside the range of the LCP in either of the suffixes, will change p.

The idea is to determine if the insertion/deletion is out of range to possibly affect the LCP value of a suffix. Recall that the LCP value for a suffix at position $i-1$ is at minimum one less than the LCP value for the previous suffix at position i, as discussed in chapter 4. We can determine that it is impossible for the LCP value at position $i-1$ to change, if neither of the LCP values which is affected by the suffix at position i changes, unless the insertion/deletion happens at position $i-1$. This is explained by the fact that at minimum, the LCP of the suffix at position i covers every character of the LCP of the suffix at position $i-1$ except for the first character. If the first character was the position of the insertion/deletion, it can change the LCP value, but that case is already handled by the positions stored in the bitset. Since the suffix at position $i-1$ did not update, this holds recursively for the next suffix at position $i-2$ as well. Therefore, as soon as we find a single suffix at position i which does not lead to an LCP value update, this will recursively hold for all suffixes at position $j < i$.

With this in mind, the algorithm to update the LCP array has two stages. Given the bitset of positions which need to be updated in correlation with the first two cases, we perform a *select* operation to continuously find the location of set bits in the bitset, and use their position to determine which positions in the LCP array need to be computed. After all the positions in the bitset has been updated in the LCP array, we move on the last two cases. From the position of pos, which was computed during the suffix array update, we continuously call the LF function on pos to find the previous suffix. At each position, we update the LCP value of the suffix at position pos and $pos + 1$, as these two LCP values are both affected by the suffix at position $SA[pos]$. When we encounter the situation where both LCP values at position pos and $pos + 1$ did not lead to a change in LCP values, the algorithm is finished and the LCP array is updated. Algorithm **??** shows how the LCP array is updated, given the bitset of updated positions which is built during algorithm 12.

The actual computation of LCP values is also fairly complex. In the incremental detection algorithm we are not storing the entire fingerprint, we only have the BWT stored in the wavelet matrix, and the LF function allows us to move from one character to the previous in the original input text. When we compute LCP values, we need to be able to compare characters in two suffixes from start to potentially end of the input, therefore we need to be able to move from one character to the next, which is the inverse of the LF function.

The inverse LF function for a position i computes the next cyclic-shift by first determining what character is in the first column of the CS at position i in the sorted cyclic-shifts, and its *rank* in the first column. With the character and rank of that character, we can perform a *select* on the BWT for that character and rank, which will give us the location of where that character is located in the last column (BWT) which is the next CS. This is the next cyclic-shift because it is the cyclic-shift where the character which was previously in the first column, but is now cycled one character to the left, which corresponds to the next CS. To find which character is in the first column at position i, recall that the F column in the sorted cyclic-shifts just contain all the characters of the input in sorted order. We can therefore use the C to determine which character is located at exactly position i, just by counting the number of occurrences of the previous characters. If C is instead implemented as an array where the $C[c]$ holds the number of characters smaller than c in the input, finding the correct character can be done efficiently with a binary search. When we know which character is located at position i in the first column, we do $select_c(i - C[c])$ to determine where that occurence of c is located in the last column (the BWT), which is then returned as the result of the inverse LF function.

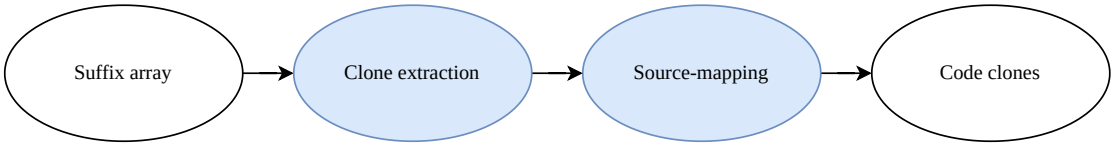
With the inverse LF function, computing the LCP value between two suffixes at position i and j is done by repeatedly applying the inverse LF function to i and j, and comparing the characters at those positions in the BWT.

TODO: Algorithm for iterating over suffixes which need to update their LCP value

TODO: Algorithm for updating a single LCP value

TODO: Visualize this somehow?

5.8 Dynamic clone extraction and source-mapping



The LCP array is no longer implemented as an array, but as a balanced tree. As trees are more costly to traverse linearly compared to arrays, we also want to improve how we extract clones in this phase. The goal of the clone extraction phase is still to find the indices in the fingerprint which corresponds to the beginning of a code clone in the original source code. These indices will be mapped back to the original source code in the source-mapping phase.

Recall that the initial detection clone extraction algorithm in chapter 4 traversed the suffixes from beginning to end, and extracts indices of suffixes where the LCP value is above the token threshold parameter. Also recall that some clones were filtered, those being clones which extend past a single fragment, or clones which are contained inside a larger clone. This algorithm will now be converted to run on the balanced tree which contains the LCP values, which also optimizes the algorithm to reduce the amount of LCP values we need to examine.

The idea for this algorithm is to always store a list of all the nodes in the tree which have an LCP value above the token threshold. These are afterall the only nodes which could possibly be the beginning of a code clone, even if some of them will be filtered out later. When a value is updated, we can check to see if the new value is above the token threshold. If the previous value was below the threshold, and the new value is above the threshold, we know that this is a node which could potentially be a code clone, so storing these nodes can reduce the amount of nodes we need to examine. When all LCP values have been updated, we can filter out nodes which have had their LCP value updated to be below the threshold.

After all LCP values are updated, we are left with a set of all the nodes with an LCP value above the token threshold. In order to filter out clones which are contained within other clones, we need to add a similar check to the loop which was added in Algorithm 8, however since the set only contains nodes with LCP values above the token threshold, the algorithm will be slightly different. Instead of adding a loop which skips past indices which are contained clones, we can instead check if the last node which was deemed the start of a code clone contains the current clone we are looking at. Doing the process in this fashion allows us to skip checking many nodes if there are long stretches of suffixes in the fingerprint with LCP values below the token threshold. Algorithm 17 shows how these clone indices are computed.

After applying this algorithm we are left with the nodes with inorder ranks which correspond to the same indices as in the initial detection For the final phase, which involves mapping the clone indices back to the original source code, we can do almost the same procedure as in the initial detection. This algorithm is similar to Algorithm 10, but some parts are changed to avoid traversing through the trees many times to find the second index, which we would have to do if we used Algorithm 10 without any modifications. The problem with Algorithm 10 is that it computes the expression SA[ISA[firstIndex] − 1] to find the index of the matching clone. This is not a problem in the initial detection, because SA and ISA are arrays, which give $O(1)$ time access. This is not the case for the dynamic extended suffix array, which would require $O(\log n)$

```
Algorithm DynamicCloneExtraction(nodesAboveThreshold)
    clones ← empty list
    lastAdded ← −1
    lastLCP ← −1
    for node in nodesAboveThreshold do
        index ← GetRankOfNode(node.pointer)
        if lastAdded ≠ −1 then
            difference ← index − lastAdded
            if node.key + difference ≤ lastLCP then
                continue
            end
        end
        Add(clones, node)
        lastAdded ← index
        lastLCP ← node.key
    end
    return clones
```

Algorithm 17: Extract clone indices in dynamic extended suffix array

access time for that expression. Instead, only the nodes of the clone indices in Algorithm 17 is stored, and for a node $firstNode$, when we can find the node of the matching $secondNode$ in $O(1)$ time by finding the predecessor of firstNode and following its pointer. Remember that the nodes we have stored as clone indices are nodes in the A tree which corresponds to SA indices. Following the pointer corresponds to SA values, or alternatively, ISA indices. After we have both the firstNode and secondNode, we can get their fingerprint indices by getting the inorder rank of their pointer, with Algorithm 15. Afterwards, the algorithm continues as previously. While this still requires traversing the tree in $O(\log n)$ time, we have reduced the number of traversals needed from 5 with a direct translation of Algorithm 10, to 2 traversals needed with the smarter traversal of nodes described here.

Now that we have successfully source-mapped from our dynamic extended suffix arrays to code clones, we are finished, and we have the set of clones ready to be sent to the client.

Chapter 6

Evaluation

In this chapter, CCDetect-LSP will be evaluated based on different criteria, which combined will provide a basis for evaluating the tool as a whole.

The most important aspect of CCDetect-LSP to evaluate is the real-time performance. We will compare the time of the initial detection with the incremental updates, as well as compare its performance with another incremental clone detector, iClones [17]. Note that we will also distinguish between the initial detection where parsing the entire code base is necessary, and subsequent detections which still constructs the suffix array from scratch, but does not require parsing the entire code base. We will call this type of detection the SACA detection, while the detection which uses the dynamic extended suffix arrays will be called the incremental detection. This is done to compare the dynamic extended suffix array against building the suffix array from scratch.

This chapter will evaluate CCDetect-LSP in the following ways:

- Verify correctness of clone detection with BigCloneBench.
- Informal complexity analysis of phases in initial, SACA and incremental detection
- Performance benchmark of SACA detection, incremental detection and iClones
- Memory usage benchmark of SACA detection, incremental detection and iClones
- Language and IDE support

6.1 Verifying clones with BigCloneBench

In order to verify that CCDetect-LSP correctly identifies clones, we should analyze and confirm that CCDetect-LSP finds all clones in a given code base. BigCloneBench [40] is a clone detection benchmark of a set of known clones in a dataset called “IJaDataset”. The benchmark consists of the IJaDataset and a database containing information of the clones which exist in the dataset. BigCloneEval [41] is a tool which simplifies evaluation of a tool on the BigCloneBench.

BigCloneEval evaluates a tool by running the tool on the dataset, letting the tool output all the clones the tool finds, and then matching the clones the tool found with the clones in the database. This makes the BigCloneBench essentially a big oracle for clone detection in a code base, which we test CCDetect-LSP against. The clones in the database are manually verified by humans, and include type-1 to type-3 clones.

We evaluated CCDetect-LSP by running the initial detection algorithm on the dataset, and outputting all the clones the tool found. BigCloneEval expects to get a file where each line contains a clone pair, where a clone is specified with filename, starting and ending line. This was simple to extract from our clone-map, where we converted our ranges which point to the beginning and ending token, to just the line number of those tokens.

BigCloneEval reports the recall of a tool, meaning the percentage of found clones. Therefore, we did not make sure that our conversion to the format BigCloneEval expects gave the minimum number of clone pairs. For example, for a clone pair of clones A and B, we output both that A is a clone of B and that B is a clone of A, which is superfluous. This would lead to a bad precision, but does not affect the recall in the report which BigCloneEval outputs.

For CCDetect-LSP, we used mostly the default parameters of BigCloneEval, but we increased the minimum clone size to 100 and set the minimum similarity threshold to be 100%, since we are only evaluating CCDetect-LSP for type-1 clone recall. The similarity threshold was set to 100% because we only consider type-1 clones which are exactly similar, and not reporting type-3 clones drastically reduces the time the evaluation takes. The default token threshold was set to 100 instead of 50, as the evaluation time increases drastically as the number of clones to match in the database and the number of clones reported by CCDetect-LSP increases.

Appendix 9.1 shows the command which runs BigCloneEval, and the beginning of the report generated when CCDetect-LSP is evaluated. The report shows that CCDetect-LSP has a ∼99.98% recall for type-1 clones. CCDetect-LSP is clearly capable of detecting type-1 clones. As for the missing type-1 clones, while it is not so simple to determine which clones were not detected, it is possible that these clones are not reported because of inconsistencies in the database. The report also shows that CCDetect-LSP manages to detect a few type-2 clones. Detecting these are accidental, as CCDetect-LSP does not currently implement any normalization of the input which would allow for type-2 detection. These type-2 clones are likely detected because BigCloneEval allows some leniency in terms of how much a clone reported by CCDetect-LSP needs to match with a clone reported in the BigCloneBench database. Therefore, it is likely the case that a few type-2 clones overlap with a type-1 clone, and BigCloneEval matched these clones because of its leniency.

6.2 Time complexity of detection

In this section we will conduct an informal analysis of the running time of each phase of the initial, SACA and incremental detection to argue that the average runtime complexity of the incremental detection is more efficient than the initial detection. In each phase we will argue the run time in terms of Big O notation. Some claims will be substantiated in the next section where we look at concrete code bases and their properties.

First, we will define some notation for this analysis. We are analyzing a code base of size n, where n is the number of characters in the code base. The fingerprint of the code base is of size

f, where f is the number of symbols in the full fingerprint and $f \ll n$. The alphabet of the fingerprint has a size of σ. In the incremental detection, E is an edit, consisting of e characters. We will also use |clones|, |documents|, |edits| when discussing the number of clones, documents and edits.

The initial detection runs in $O(n)$ time. The bottleneck of the initial detection is reading and parsing all the content in each file. Tree-sitter generates Generalized LR (GLR) parsers [29], which in the worst-case have a $O(n^3)$ running time, but $O(n)$ for any deterministic grammar. As programming languages generally have deterministic grammars, we will assume that the running time of parsing with Tree-sitter takes $O(n)$ time. After the initial parsing, the SACA detection runs in $O(f)$ time. The running time is $O(f)$ because the suffix array construction is performed for every update, which takes linear time in the size of the input, which is the fingerprint. The extraction of clones from the LCP array also runs in $O(f)$ as it is a single scan over the LCP array. The final source-mapping is a bit more complicated, taking $O(|\text{clones}| \times \log(|\text{documents}|))$. This complexity comes from binary-searching the list of documents to find the correct document for each clone. This is highly likely to be less time consuming than the suffix array construction, as the number of documents and number of clones are usually orders of magnitude lower than the size of the whole code base. Therefore, we get a final running time of $O(f) + O(|\text{clones}| \times \log(|\text{document}|))$, where $O(f)$ is highly likely to be the dominating factor.

For the incremental detection, we have already parsed the code base and built the index and dynamic extended suffix array structure for the code base. Afterwards, when an edit E is performed in a document D, the first phase is to update the document index. We iterate over the documents to update their fingerprint ranges, which has a complexity of $O(|\text{documents}|)$. When a document which has been edited is reached, the new document content is incrementally parsed, queried for fragments and then the fragments are fingerprinted. In an IDE scenario, only one file is edited in each update, so only one document needs to be parsed in each update. In the worst-case, we have to fingerprint the entire document, which has a complexity of $O(|D|)$. The complexity of this phase is therefore $O(|\text{documents}| + |D|)$.

Next, extracting the edit operations which have happened to D_f (fingerprint of D), takes $O(|D_f| + |E|^2)$ where $|E| \leq |D_f|$. Note that the size of the edit is calculated as the area which E covers, meaning that if an edit consists of changing a token at the beginning of the file, and a token at the end of the file, then $|E| \approx |D_f|$. We get this time complexity because Hirschberg's algorithm runs in $O(n \times m)$ where in our case, $n \approx m$. If the size of the edit is contained in a smaller area, we apply the optimization which reduces the size of the edit by comparing characters at the beginning and end of the string, as discussed in chapter 5. This process has a complexity of $O(|D|)$, and afterwards Hirschberg's algorithm has a worst-case complexity of $O(|E|^2)$.

The worst-case complexity of dynamically updating the extended suffix array is actually slower than a linear time SACA algorithm in the worst-case. The worst-case scenario when inserting/deleting a character in the fingerprint is that every single suffix needs to be reordered, meaning we reorder $O(f)$ suffixes, where each reordering takes $O(\log(f))$ time, as it requires deleting and inserting an element in the dynamic extended suffix array. In addition, we need to call the LF function twice for each reordering, which has a $O(\sigma + \log f \log \sigma)$ complexity. The complexity of the LF function comes from a *rank* call in the wavelet matrix, and an iteration over C to find the number of smaller characters. This results in a $O(f \times (\sigma + \log f \log \sigma))$ running time of this phase, which is worse than the $O(f)$ running time of the SACA algorithm. In addition, before the reordering when we are inserting/deleting characters in the BWT, we use the LF function and

insert/delete in the wavelet matrix and the C array. The dominating factor of these operations is again the LF function. If an edit operation consists of inserting/deleting multiple characters, the number of LF function calls is increased, with a complexity of $O(e \times (\sigma + \log f \log \sigma))$. Note that while there are many factors in this time complexity, all the factors should be quite small compared to f, making the complexity of reordering the suffixes the dominating factor of this phase.

The average-case complexity of this phase is however highly likely to be faster. Salson et al.have shown that on average, the number of reorderings required for an insertion/deletion in a suffix array is highly correlated with the average LCP value of the input [30]. Their data shows that for multiple different types of data such as genome sequences and english text, the average LCP value of the input is generally magnitudes lower than the input size. In our experiments, this applies to source code as well, as code bases we have tested on have all had an average LCP values well below 100. See Table 6.1 to see the average LCP value for different code bases. A lower number of reordered suffixes for lower LCP values seems intuitive, as lower LCP values mean that the insertion/deletion will affect the ordering of fewer suffixes in the input. With this information, it would be more accurate to de-emphasize the importance of the $O(f)$ number of reordered suffixes in our analysis, and we therefore claim that the average running time of an edit operation on the suffix array is closer to $O(e \times (\sigma + \log f \log \sigma))$ if we account for multiple characters being inserted or deleted as well. We extend this to account for multiple edit operations as well, for each edit operations which was computed in the last phase, we perform an insertion/deletion of e characters. Therefore, the total complexity of this phase on average is closer to $O(|\text{edits}| \times (e \times (\sigma + \log f \log \sigma)))$.

Next there is the clone extraction phase. Recall that in the dynamic detection clone extraction phase, we had stored all nodes with an LCP value above the token threshold, and iterate over those to determine which of them are clones or not. In the worst-case, every index in the LCP array would be above the token threshold, which would be an $O(f)$ complexity iteration. However, this is highly unlikely, since the nodes with LCP value above the token threshold are either clones, or contained clones. The number of contained clones is limited by the number of clones because each clone can only have a limited amount of contained clones. Therefore, the number of LCP nodes examined should be closer to $O(|\text{clones}|)$, which is highly likely to be much less expensive than iterating over all the LCP nodes. For each examined LCP node, we need to traverse from the node, to its pointer node, and then to the root of the B tree to determine the fingerprint index of that LCP node. We do the same to determine the index of the matching clone, as described in Chapter 5. These traversals take $O(\log f)$ time. The worst-case performance of this phase is therefore $O(f \times \log(f))$, but we will see that the average-case performance is closer to $O(|\text{clones}| \times \log(f))$.

Finally, the source-mapping phase is easier to analyze. As we now know all the positions of clones and their matches, building the clone-map takes $O(|\text{clones}| \times \log(|\text{documents}|))$ time. We perform the binary-search over the documents to get the source location for each clone index we found in the previous phase. This is the same complexity for both the SACA and incremental detection.

With this informal analysis, we have demonstrated that in the average case, the incremental detection should be faster than the SACA detection, as none of the phases reach or exceed the $O(f)$ running time of the SACA detection. In the next section we will show that the properties we have assumed for this analysis are present in multiple code bases, and that these properties lead to a better benchmark performance for the incremental detection.

6.3 Benchmark performance

For the performance benchmark we will benchmark the running time of CCDetect-LSP on code bases and edits of different sizes. We will compare the performance of the incremental detection with the SACA detection and iClones [17].

iClones is a tool which, similarly to CCDetect-LSP, does incremental clone detection. The tool takes n revisions of the same code base and will after the initial detection, reuse as much information as possible to reduce the amount of work needed to analyze consecutive revisions. The algorithm which iClones uses is based on generalized suffix trees (GST), and the tool can detect up to type-3 clones. For a revision i, iClones expects to have a file named `changed` in the root folder of the revision, which contains information on which files have changed between revision $i-1$ and i. For each file which has changed, the suffixes in the file is inserted into the GST, reusing nodes of the tree if possible.

For the performance benchmark, the code bases are set up in the way iClones expects. For each code base to analyze, we will store n revisions of it, where each revision contains some changes to some files. An evaluation program for CCDetect-LSP was written so that it can process this same setup, making it simple to run both CCDetect-LSP and iClones on the same code bases, and compare their results.

To create a larger sample set of code bases and edits, it is preferrable to generate new revisions of a code base programmatically. Generating new revisions of a code base means that we copy the code base and apply a number of insertions/deletions to random files, to simulate how a programmer would edit files when programming. Generating new revisions of code bases is not a trivial problem, since each edit performed in a file should still give a valid program which can be parsed. If the file cannot be parsed, Tree-sitter will not give a proper AST for the file. Because of this, generating insertions is hard, because we need to make sure that the location we are inserting some code is a valid location for that code. It is easier to generate deletions first instead. Deletions are easier to generate, as we can use Tree-sitter to determine where a fragment of size n is located, and remove that section of code in the next revision. For example in Java, we can delete an entire method, and still know that the program will parse to a valid AST afterwards. After we have generated for example 10 revisions where the difference between each revision is deletion of some number of fragments, we can then reverse the ordering of the revisions, so that the final revision is now the first revision, and so on. This will translate each deletion to instead be an insertion, as an insertion is simply an inversion of a deletion. Figure 6.1 illustrates how a test consisting of three revisions with deletions can be inverted to create a test of insertions. With this technique we can generate arbitrary tests consisting of either a number of insertions or deletions in each revision.

There are multiple variables we can tune when generating the tests. Primarily the three most important factors of the tests is the size of the code base, the number of edit operations in each revision, and the size of those edit operations.

The size of the code base will naturally affect the running time of the algorithm. We randomly picked open-source code bases of differing sizes and differing amounts of duplication to run the

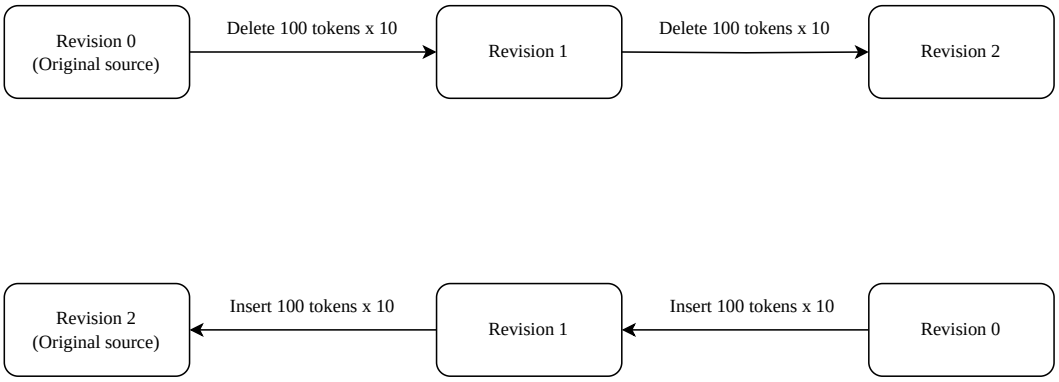


Figure 6.1: Reversion of an evaluation test. Delete100x10 to Insert100x10

Code base	**LOC**	**Clones detected**	$\mathbf{LCP_{avg}}$	$\mathbf{LCP_{\geq 100}}$
WorldWind	550KLOC	1517	18	63967
neo4j	1MLOC	1313	9	27557
graal	2.2MLOC	2012	28	154452
flink	2.3MLOC	4729	13	155754
elasticsearch	3.2MLOC	9986	14	289511
intellij-community	5.8MLOC	3585	19	336190

Table 6.1: Properties of code bases

benchmark on. We selected the code bases WorldWind[1], neo4j[2], graal[3], flink[4], elasticsearch[5] and intellij-community[6]. The code bases are all Java code bases of different size, with different amounts of duplication. Table 6.1 shows some properties of the code bases, including its number of lines, the number of clones detected by CCDetect-LSP for a token threshold of 100, the average LCP value and the number of LCP values above 100. An interesting aspect of the graal and flink code bases is that they both contain ∼2.2MLOC, but graal has a higher LCP_{avg}, which could potentially lead to a slower suffix array update.

Additionally, we chose to test both an edit size of 10 and 100 with 10 edits in each revision. We believe these values to be slightly larger than what is realistic for an IDE scenario where a programmer is editing and updating the fingerprint of a single file at a time. For a test where we are performing 10 insertions of 100 tokens, this means we are inserting 1000 tokens in a single edit, which is likely larger than any realistic edit, but could occur in large-scale automatic refactoring operations. Therefore, if the data shows that the algorithm can process these types of edits in "real-time", we believe that the algorithm can process most realistic editing scenarios.

The benchmarks were run on a computer with 16GB RAM, and an Intel i7-2600K CPU with a 3.4GHz clock speed and 4 cores[7]. The computer runs Manjaro Linux with kernel version 6.1.25-1.

[1]WorldWind: `https://github.com/NASAWorldWind/WorldWindJava`
[2]neo4j: `https://github.com/neo4j/neo4j`
[3]graal: `https://github.com/oracle/graal`
[4]flink: `https://github.com/apache/flink`
[5]elasticsearch: `https://github.com/elastic/elasticsearch`
[6]intellij-community: `https://github.com/JetBrains/intellij-community`
[7]Note that this CPU is quite old, and may not be representative of today's modern CPUs

The SACA detection, incremental detection and iClones was run with a token threshold of 100, and the processing time of each revision was timed using a simple clock mechanism programmed in Java. Type-3 clone detection was disabled for iClones with the `-minclone 100` parameter. Each test was run three times, and the results were averaged. Only three runs of all tests were performed because running all the tests takes a significant amount of time (multiple hours), and the results seem to be stable from three runs only.

In addition, we performed a final test on the elasticsearch code base where in each revision, the number of edits increase. This was tested to see how CCDetect-LSP scales when the number of edits increase. The size of these edits are not realistic for the IDE scenario, as when the number of edits increases, thousands of tokens are being edited in a single revision. However, this could be more realistic for the evolution scenario. Recall that the evolution scenario is when we analyze different revisions of a code base such as different commits in version control.

Figure 6.2, 6.3, 6.4, 6.5, 6.6, 6.7 and 6.8 shows the benchmark results on the different code bases. Each graph shows the running time of the SACA detection, the incremental detection and iClones when run on a certain code base and edit operation size. The running time is shown on a log scale, as the initial detection is extremely slow compared to the subsequent detections. iClones was not able to run on the intellij-community code base, as the memory usage exceeded 16GB, therefore only the CCDetect-LSP algorithms is shown on those graphs. The elasticsearch code base with a size of 3.2MLOC was approximately the maximum size code base iClones could run on before the computer ran out of available memory. Note that figure 6.8 shows the test where an increasing number of edits were performed in each revision, the number of edits are displayed as the labels of the X-axis.

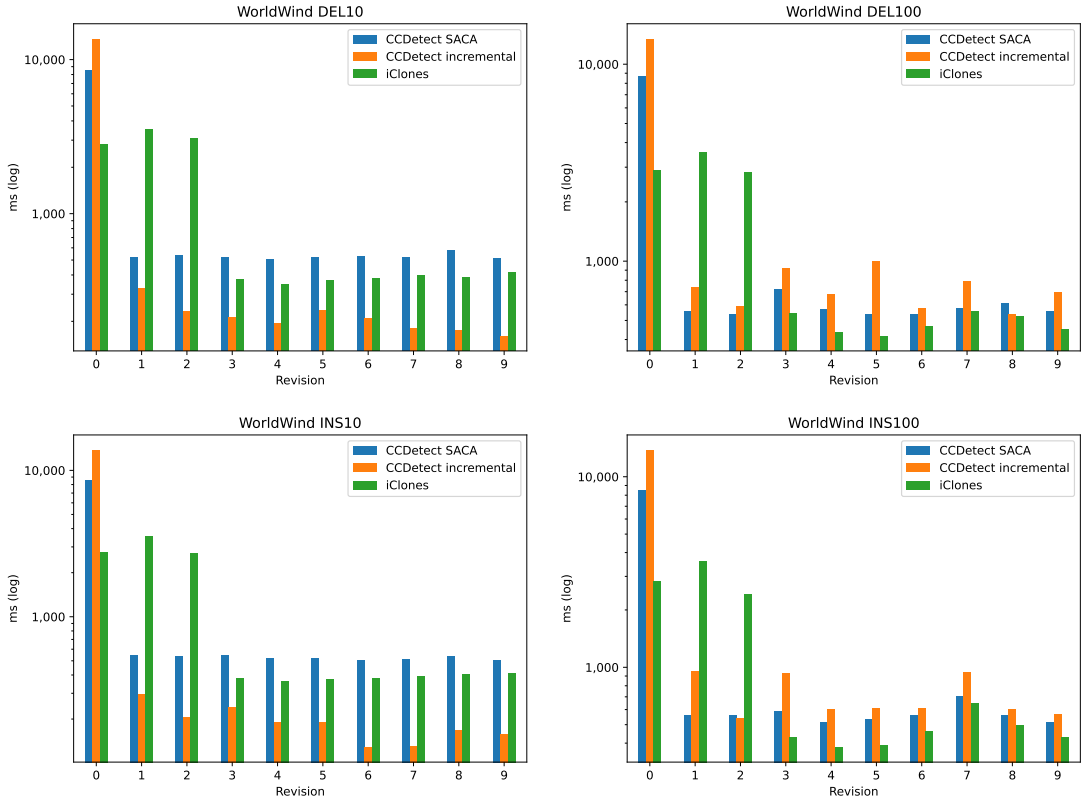


Figure 6.2: WorldWind performance benchmark

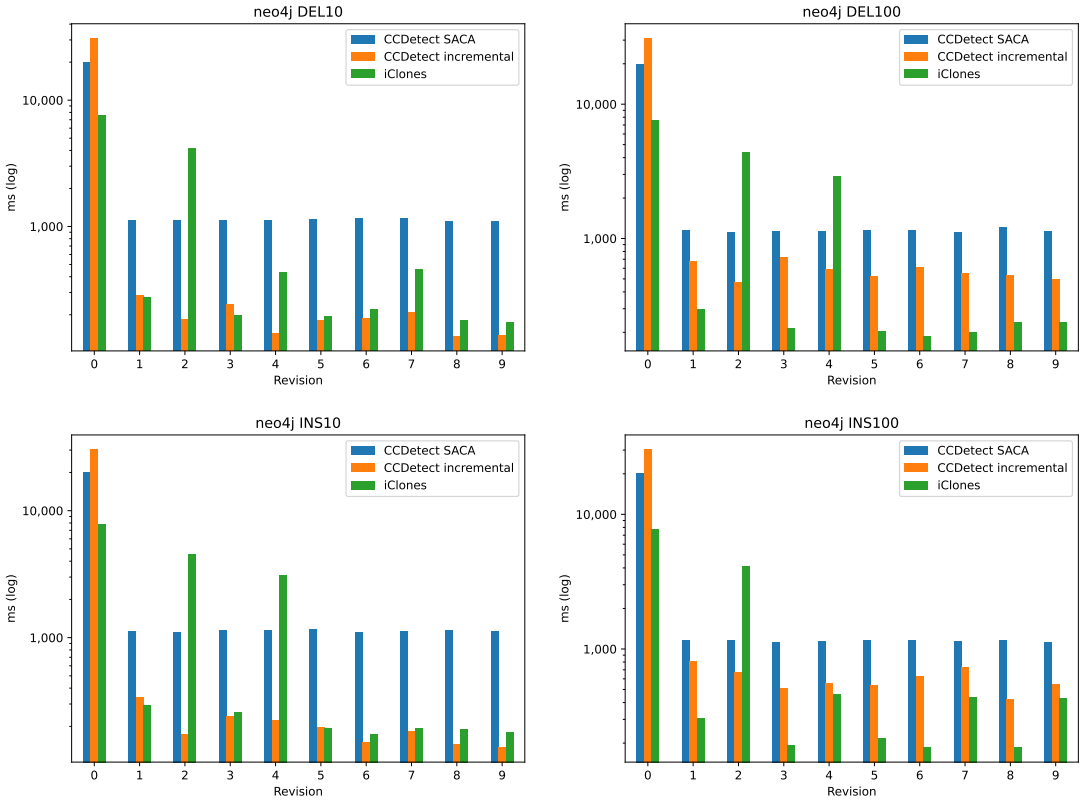


Figure 6.3: neo4j performance benchmark

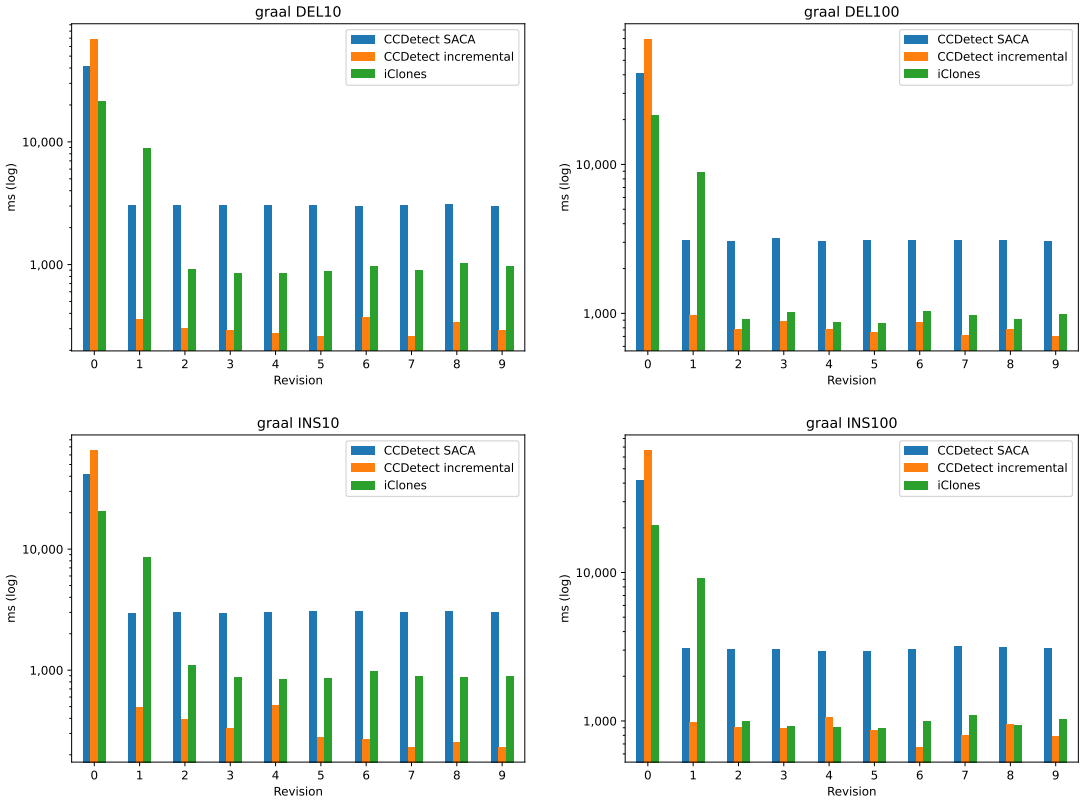


Figure 6.4: graal performance benchmark

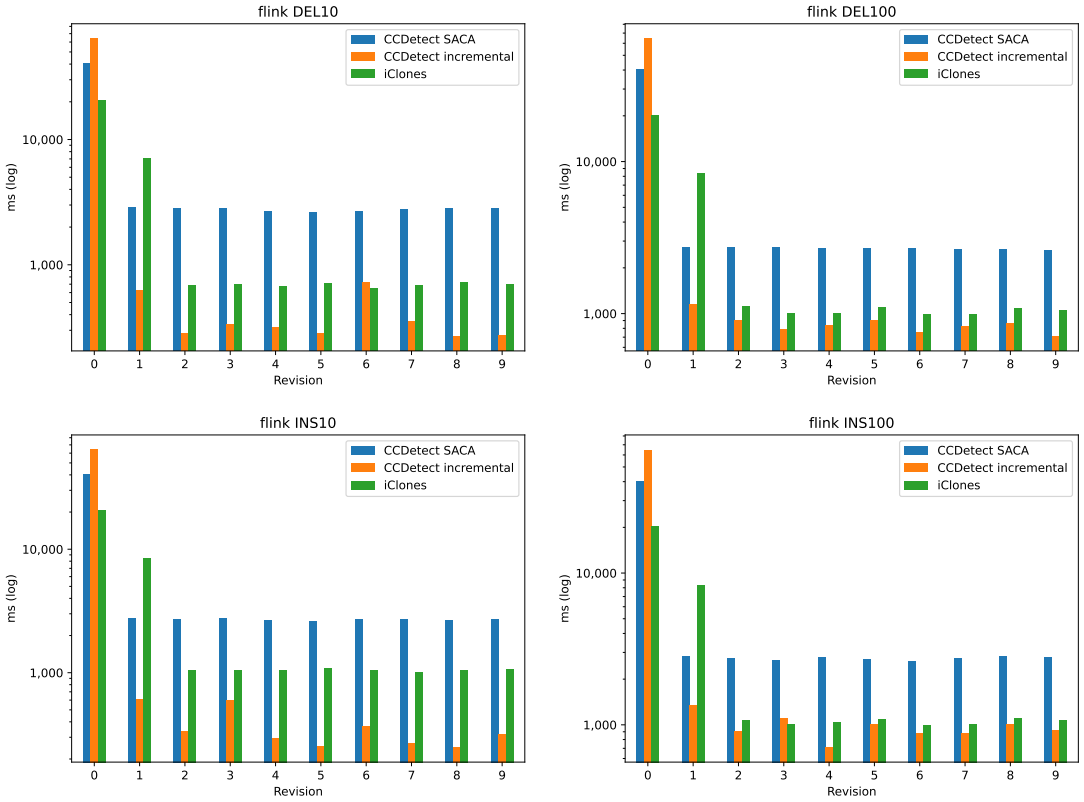


Figure 6.5: flink performance benchmark

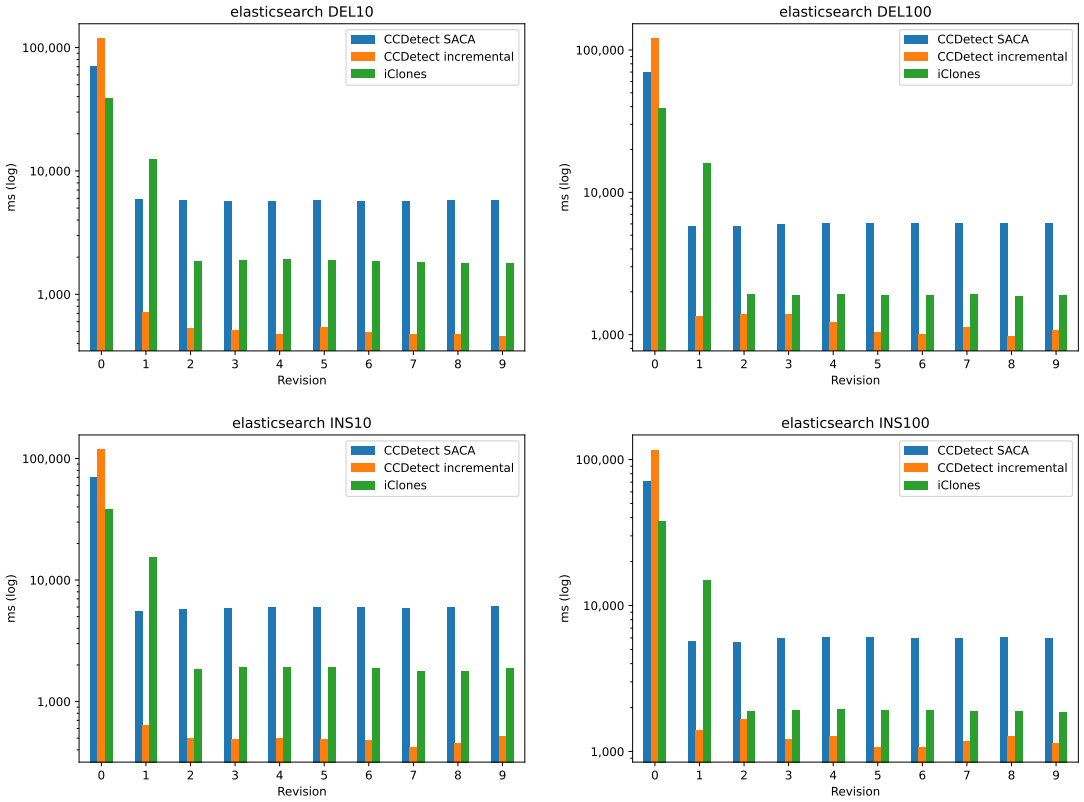


Figure 6.6: elasticsearch performance benchmark

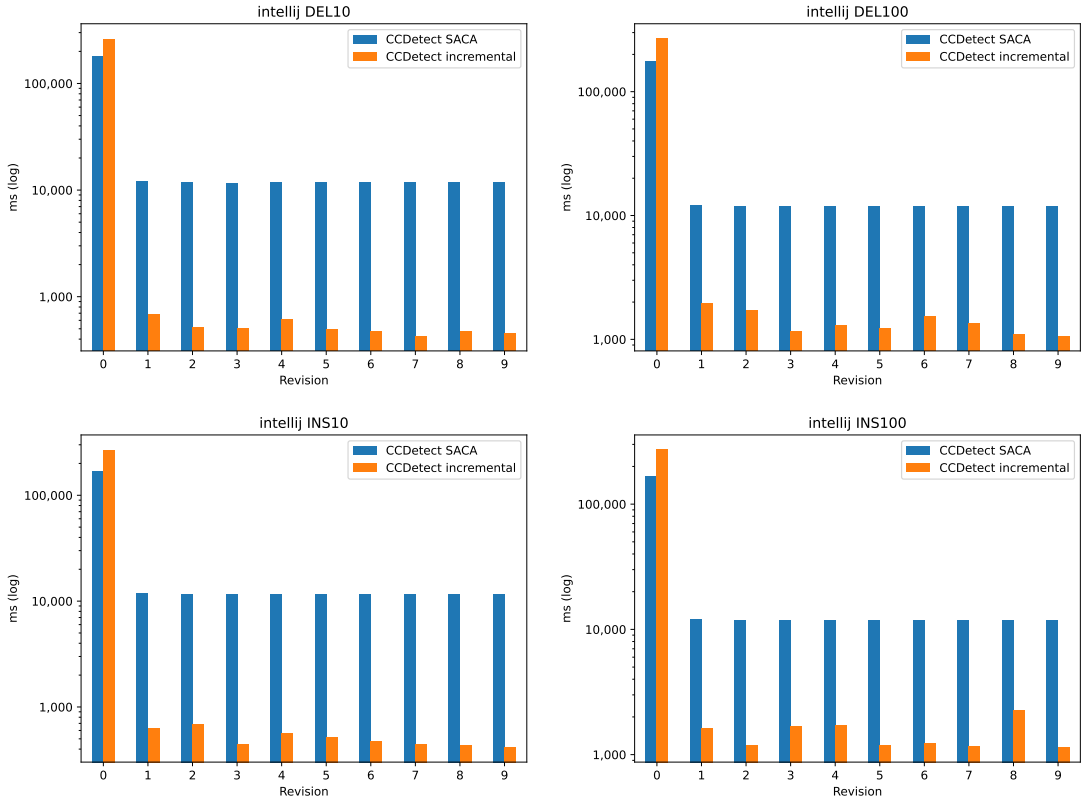


Figure 6.7: intellij-community performance benchmark

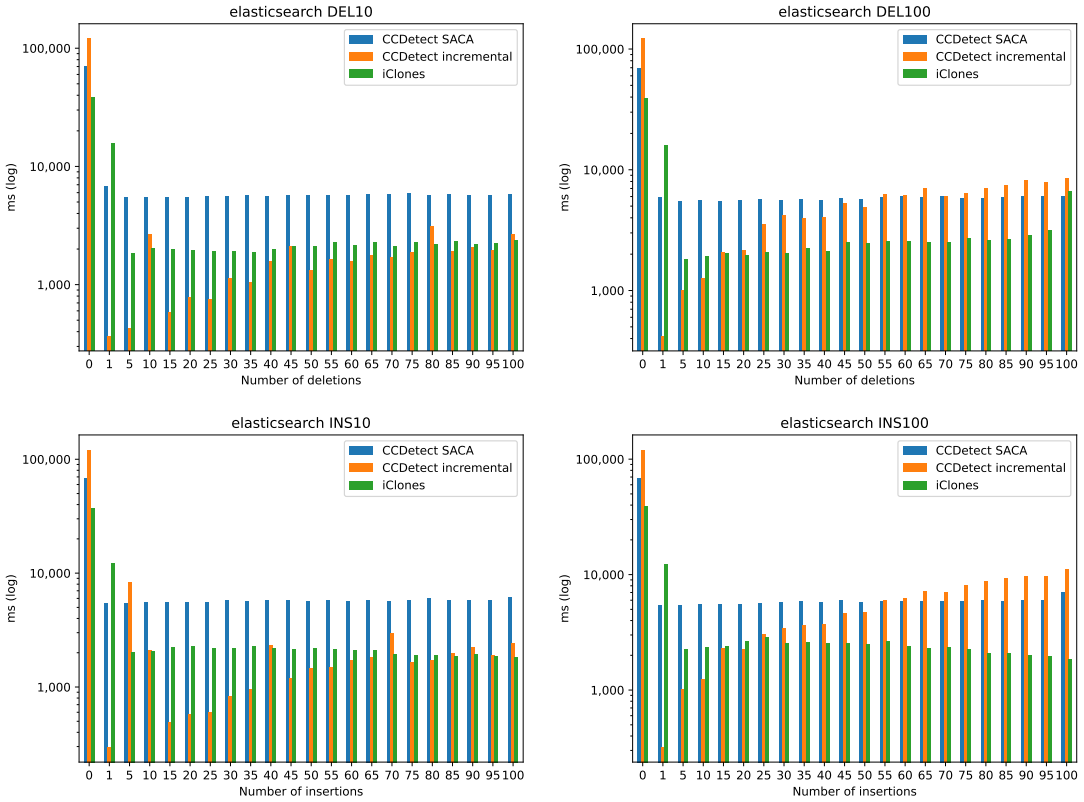


Figure 6.8: Elasticsearch performance benchmark with increasing number of edits

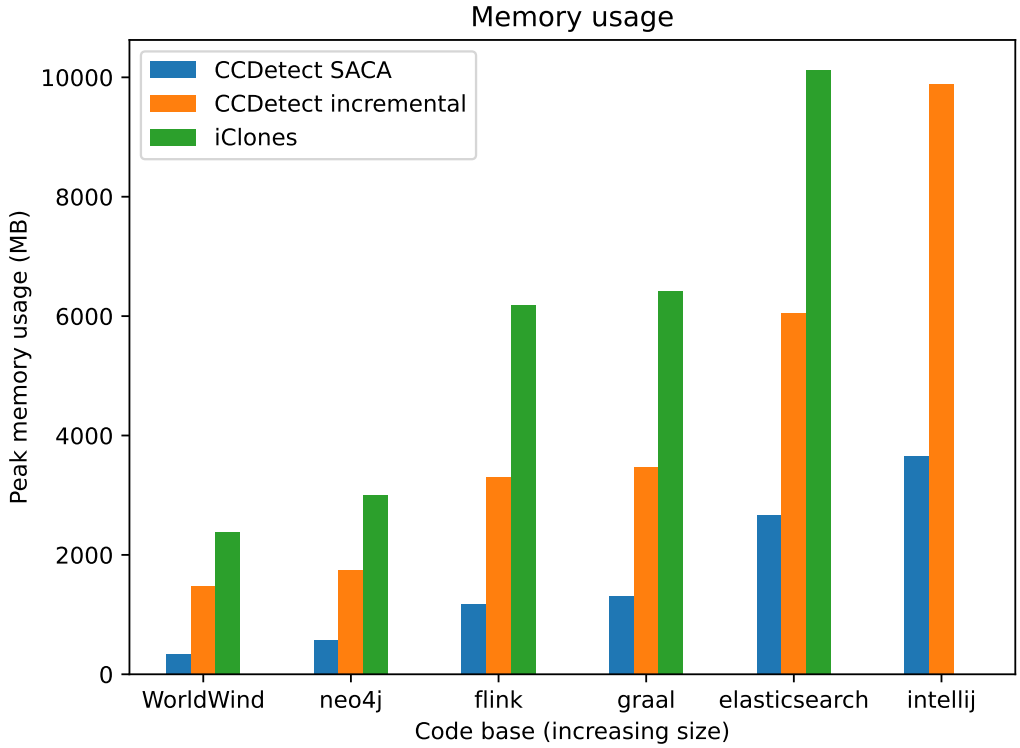


Figure 6.9: Memory usage of each tool when running the DEL10 test for each code base.

6.4 Memory usage

Another important aspect of clone detection tools and especially IDE tools, is the memory usage of the tool. As CCDetect-LSP is primarily implemented as an IDE tool, it is important that the tool can run alongside other IDE tools and applications. Therefore, we will in this section analyze the memory footprint of CCDetect-LSP and again compare the SACA detection, incremental detection and iClones. To simplify this analysis, we will for each code base run one of the tests we previously ran for the performance benchmark, and observe the peak memory usage of each algorithm, using the JProfiler[8] profiling tool. In our experience, the clone detection tools we are evaluating in this scenario does not have large spikes in memory usage, and generally stabilizes around the peak memory usage. Therefore, we only observed the peak memory usage, as it seems to represent the amount of memory which would be required to run each tool in an IDE scenario.

JProfiler provides a live memory usage overview which it records every 2 seconds on a running JVM process. Figure 6.9 shows the results of the memory usage test. iClones is again left out for the intellij code base, as the memory usage exceeded 16GB.

[8]JProfiler: https://www.ej-technologies.com/products/jprofiler/overview.html

6.5 Languages and IDEs

CCDetect-LSP attempts to achieve the goals of being IDE agnostic by utilizing LSP, and programming language agnostic by utilizing Tree-sitter. In this section we will quickly go over our experience with using these tools, and what we have implemented and tested to determine if these goals are met.

Language agnostic

For the goal of being language agnostic, we have tried CCDetect-LSP with multiple languages, including Java, Python, C, Rust, JavaScript, and Go. To add a new language to CCDetect-LSP all that is needed to be done is to download the correct grammar into the `grammars/` directory of the CCDetect-LSP source code, and add one case to a match-statement in order for CCDetect-LSP to tie that grammar to a certain file-type. From there, the client needs to change its configuration to the new language. This involves only changing two configuration parameters, the file-type to analyze, and the Tree-sitter query to be used in the fragment selection phase. Determining the Tree-sitter query may not be trivial in some languages, but we suspect that the most common choices would be either selecting the root node of the AST, or all methods or functions[9].

Initial experiments show that CCDetect-LSP seems to work flawlessly on most languages. There experiments are not shown in the thesis and are based purely on our experience with testing the tool on different code bases. While we cannot objectively tell if all clones are detected without having an oracle like the BigCloneBench for each code base, the experience of browsing and navigating clones seems to work as well when switching between different languages. For example, we explored the source code of both the Rust and Go compilers in their respective languages, and we were able to locate thousands of clones in each code base (token threshold 100).

While our testing of multiple languages was not extensive, only one language we tested did not work flawlessly out of the box. The Tree-sitter grammar for C had a weird quirk where not every token had its own leaf-node in the AST. For example, the string literal token did not have a distinct leaf-node for the string value. Therefore, functions which had completely different string literals in them could be considered clones. In a positive light, this actually resulted in some type-2 clones being detected, but it also resulted in some clones being detected which were not clones at all. Even if this issue was preventable by treating some AST nodes differently, we consider this to be an instance where CCDetect-LSP did not have perfect language support.

IDE agnostic

CCDetect-LSP can in theory communicate with any IDE which can act as an LSP client. We have tested CCDetect-LSP on two IDEs which have great LSP support, Neovim[10] and VSCode[11].

For Neovim, only a configuration file needs to be setup, where we specify the command which launches the LSP server, which file-type the LSP server should be launched on, and the rest of the configuration which CCDetect-LSP needs. Diagnostics and code-actions works out of the box, but a diagnostic view which shows all code clones requires an extra plugin such as

[9] Tree-sitter queries for multiple languages are shown in the README file of the CCDetect-LSP source code.
[10] Neovim: `https://neovim.io/`
[11] VSCode: `https://code.visualstudio.com/`

Telescope[12]. Neovim does not support the `DiagnosticRelatedInformation` information which can be attached to diagnostics, therefore listing matches and navigating to them is achieved with code-actions.

For VSCode, a small extension (plugin) had to be created, which simply uses the built-in APIs of VSCode to launch the LSP server on the correct file-type with the correct configuration options. Once launched, VSCode has full LSP support, including diagnostics with match information, code-actions and a nice diagnostic view which displays all code clones. VSCode running this plugin was shown in figure 3.1 in chapter 3. While the configuration settings such as token threshold and language is currently hard-coded into the extension, we imagine it would be simple to create a GUI configurator for CCDetect-LSP in VSCode. It could also be possible to use the `didChangeConfiguration` message defined by LSP to change the configuration as the LSP server is running, which would allow changing for example the clone token threshold without restarting the LSP server.

[12]Telescope: `https://github.com/nvim-telescope/telescope.nvim`

Chapter 7

Discussion

This chapter discusses the results of the evaluations performed in chapter 6 and conclude with some insights into which algorithms seem to be more efficient in each case. We will also discuss the implications of the memory usage, and discuss the clones which are detected, and the possibility of extending the incremental algorithm of CCDetect-LSP to more intricate types of clones.

7.1 Performance

In chapter 6 we saw how the SACA detection, incremental detection and iClones performed on multiple code bases of differing size. This section will discuss these observations and consider which factors that contributed to the observed outcomes and what this means for CCDetect-LSP as a clone detection tool in an IDE environment. While we have tested iClones in our evaluation, we do not have access to the source code. Therefore, we cannot speak much to why iClones is faster/slower on certain test instances. It is possible that iClones spends some time to build data structures for detecting type-3 clones, which would be unfair for the evaluation as type-3 detection is not considered. However, it seems most of the work is avoided when type-3 clone detection is disabled, as enabling it drastically slows down the benchmarks of iClones.

It is clear that for the initial detection (revision 0), iClones has the best performance. This was unexpected, as one would think that the SACA detection, which does not need to build any significant data structures, would be able to detect the clones faster than iClones. However, we also see that the subsequent SACA detections (revisions 1-9) after the initial detection is generally much faster than the initial iClones detection (revision 0). Since much of the initial detection time is taken up by parsing the entire code base, it is possible that the parsing in the SACA detection is slower than iClones parsing, not the clone detection. Tree-sitter focuses on being performant on incremental parsing, and therefore might not be as performant on the initial parse, which could explain why the initial detection for both the CCDetect-LSP detections are slower than iClones[1]. In the case of our incremental algorithm, it is not unexpected that the initial detection is slower than the SACA initial detection, and therefore also the slowest algorithm overall for the initial detection. Because our incremental algorithm does the same work as the

[1]Tree-sitter is not widely discussed in the literature, but is open-source. This claim is based on the source code and discussions surrounding it, e.g. Github issues.

initial detection, but also needs to build the wavelet-matrix and dynamic extended suffix array data structures, it seems to consistently be the slowest algorithm in the initial detection. This is one of the tradeoffs of our incremental algorithm, as the initial detection can be very slow for larger code bases, taking 3-4 minutes for the intellij-community code base. An interesting idea is to persist some of the data on the disk, in order to avoid the slow initial detection of CCDetect-LSP. If the fingerprint of each document was stored in a file on disk, this file could be read and used to index the project without needing to parse every file before the detection can happen. However, we encounter a problem whenever we want to perform an incremental update, as we do not know which tokens map to which integer values in the fingerprint. This could be solved by also persisting the token to integer mapping.

For the subsequent incremental detections (revisions 1-9), we see that depending on the size of the code base, and the size of the edits, our incremental detection and iClones are generally much faster than the SACA detection, as expected. For the smallest code bases (WorldWind and neo4j), our incremental detection seems to be the fastest for smaller edits (INS10, DEL10), while iClones seems to be faster for the larger edits (INS100, DEL100). It seems that when the code bases increase in size, our incremental detection scales better than iClones and the SACA detection. The incremental detection is able to process updates to intellij-community (5.8MLOC) in less than 1 second for smaller edits (INS10, DEL10) and less than 2 seconds for the larger edits (INS100, DEL100). Elasticsearch (3.2MLOC) was the largest code base that iClones could be run on with 16GB RAM, and for this code base it seems that our incremental detection starts to outperform the other two detection algorithms in all cases. It would be interesting to see if this trend continues for even larger code bases and continue to compare CCDetect-LSP incremental detection with iClones, but this would require more RAM to benchmark.

In the case of figure 6.8, where we performed an evaluation of elasticsearch and incresed the number of edits in each revision, we see that for a smaller number of edits, our incremental detection has the best performance. For the smaller edits (INS10, DEL10), the incremental detection outperforms or is comparable to the other algorithms even up to 100 edits (1000 tokens inserted or deleted in total). However, we see a clear trend that once the number of edits increase, our incremental algorithm is slowly overtaken by iClones. For the larger edits (INS100, DEL100), we see that our incremental detection is fastest for the smaller number of edits, but is quickly outperformed by iClones and even the SACA detection at around 25 edits (2500 tokens inserted or deleted in total). These results are an indication that CCDetect-LSP incremental detection scales worse than the other algorithms when the number of edits increase. Our incremental detection can outperform the other algorithms up to a decently large number of edits, but if the code base drastically changes (thousands of tokens) in each revision, iClones seems to have the best performance. Since editing thousands of tokens at a time is unrealistic in the IDE scenario, we are therefore confident in saying that CCDetect incremental detection still has the best performance for the IDE scenario, but that this algorithm is not very suitable in the evolution scenario, if the changes in each revision are affecting thousands of tokens.

In terms of stability of performance, it seems that the SACA detection is naturally very stable between each revision, as it is recomputing the entire suffix array anyway. The two incremental algorithms are more unstable in terms of running time, and this stems from the fact that some incremental updates are less computationally heavy than others. We cannot speak to the implementation of iClones, however, we often see that iClones takes an extra revision or two before it stabilizes at a lower running time. We can also sometimes see spikes in performance for iClones (such as in figure 6.3). Our incremental algorithm has a running time which fluctuates mostly

by how much time updating the suffix array takes. The other phases are generally either very stable (source-mapping) or are too fast to differentiate between revisions (parsing, fingerprinting). The time to update the suffix array is highly dependent on how many suffixes needs to be reordered. The number of suffixes being reordered directly affects how many LCP values need to be updated, which is often the bottleneck of the suffix array update. In our testing we examined two different code bases of size ∼2.2MLOC, graal and flink. We hypothesized that graal would have slower updates, as its LCP_{avg} was more than double that of flink. According to Salson et al. [30], this should on average lead to more suffixes being reordered. However, our results show that the running time for updates in these code bases are very similar. The reason why we do not see much of a difference could be because the LCP_{avg} differences in our code bases is too small to see any noticeable effect by doubling the LCP_{avg}. Inspecting the numbers by Salson et al. [30], shows that the number of suffixes needing to be reordered for lower LCP_{avg} values were often below 20 and that the number of reordered suffixes only drastically increases for LCP_{avg} in the hundreds. These values are more realistic for inputs such as DNA sequences than for our source code fingerprints. It therefore seems that source code is a good candidate to apply the dynamic suffix array update algorithm to, as updates in source code will on average mean few suffixes need to be reordered.

With these results, it is clear that CCDetect-LSP incremental detection is fast enough to be used in the IDE scenario for many code bases at least up to ∼2.2MLOC (flink, graal), as most updates can be processed in under 1 second for these code bases. For code bases of larger sizes (elasticsearch, intellij), small edits (INS10, DEL10) can still be processed in under 1 second, but larger edits (INS100, DEL100) can start to take some time, as the processing time exceeds 1 second and can be closer to 2 seconds. If the programmer is generally only performing small edits in few files, and not doing large-scale refactoring across many files at once, CCDetect-LSP is likely fast enough to feel "real-time" even for these large code bases, but might feel sluggish to update on such larger refactoring operations. It is also clear that while CCDetect-LSP incremental detection is fast for the IDE scenario, it starts to get outperformed by other algorithms in the evolution scenario if the number of edits increases, and the size of the edits get larger.

7.2 Memory usage

In chapter 6, we also saw the results of the memory usage evaluation for the different tools. This section will discuss these results and consider which factors increase the memory usage of each tool, and what this means for CCDetect-LSP when used in an IDE environment.

The results show a clear picture on how the memory usage scales with the number of lines of code. The SACA detection has the lowest peak memory usage, which is expected, since it only needs to store the extended suffix array in array form, which is more memory efficient than the dynamic structures of the incremental algorithms. The main memory usage for the SACA detection is the extended suffix array, which is not much larger than the source-mapping information.

For the incremental algorithms it is clear that CCDetect-LSP incremental detection has a lower memory usage than iClones, but the memory usage is still quite high compared to the SACA detection. Inspecting the JProfiler memory overview shows that the main memory bottleneck of CCDetect-LSP incremental detection is the wavelet-matrix, and the dynamic extended suffix array data structures. These two data structures take up about the same amount of memory, and since they both are pointer-based structures representing values for the entire fingerprint, these two take up a lot of memory combined. For iClones, we see that the memory usage is even more

severe, seemingly doubling the memory usage of CCDetect-LSP incremental detection. JProfiler reports that it is the suffix tree data structure which is the memory bottleneck in iClones.

While our incremental detection manages to lower its memory usage compared to iClones, the memory usage is still quite high for an IDE tool. While the dynamic structure for storing the extended suffix array gave us a large performance gain in terms of time, it is clear that it is not nearly as memory efficient as the normal suffix array in array form. For a larger code base such as elasticsearch and intellij-community, one would likely require a computer with at least 16GB RAM in order to justify running CCDetect-LSP incremental detection in the IDE. Again, it seems like CCDetect-LSP incremental detection is a good fit for code bases up to $\sim 2.2 MLOC$ (graal, flink), where the memory usage is lower than 4GB.

7.3 Clones detected

In chapter 6 we saw that CCDetect-LSP identifies $\sim$99.98% of type-1 clones in the BigCloneBench dataset. The results showed that CCDetect-LSP was able to identify practically all type-1 clones, which gives us confidence that our algorithm is correct in terms of the reported clones. BigCloneBench also reported that CCDetect-LSP was able to detect some type-2 clones, but as previously stated, this is accidental.

We decided to store and represent clones as clone classes, rather than clone pairs. We chose this because it gives a clearer picture to the programmer of all the clones of a certain code snippet, rather than having to navigate multiple times between multiple clones to determine all the clones of the snippet. Baars et al. also claims that storing clone classes is advantageous in a refactoring scenario, which is intuitive, as refactoring a set of clones likely leads to a better result than refactoring only a single clone pair [5].

Detecting type-1 clones is likely the most important clones to detect in an IDE scenario, but type-2 clones are also likely useful. Recall that type-2 clones are clones which are structurally identical, but allows differences in literals, identifiers and types. Recall also that both type-1 and type-2 clones are clones which are often good targets for refactoring, since they can be parameterized to account for the differences in literals and types, before they are merged into a single function/method. In the case of CCDetect-LSP, we could normalize the input to allow detection of type-2 clones by consistently fingerprinting tokens of the same type, but different value to the same value if they are allowed to. Because CCDetect-LSP is language agnostic and relies only on a Tree-sitter grammar for a given programming language, performing this normalization is more challenging than in typical clone detectors, and would likely require the client to send a list of token types which should be normalized. One might also want to differentiate between consistently renaming tokens and consistently substituting specific token values. An example of this is that one might only consider two snippets type-2 clones if their variable names are consistently renamed, meaning that there needs to be a consistent mapping between variable names in the two snippets. Baker's technique could be implemented to achieve this, but again, this seems difficult to achieve in a language agnostic setting [6]. This could be done by allowing the user to configure a list of AST node types which should be consistently fingerprinted, but this also requires the user to be intimately familiar with the node types in the AST of their chosen programming language.

For type-3 clones we need to consider how useful it is to report these clones in an IDE scenario. Recall that type-3 clones are clones which in addition to type-1 and type-2 clones, allow some

leniency in terms of the code clones structure. Type-3 clones allows tokens to be added, modified or removed and two type-3 clones are considered equal based on some similarity threshold. Seeing such clones in the IDE could possibly lead to a lot of noise, and CCDetect-LSP could possibly detect a lot of clones which are not necessarily simple to refactor and remove. It would be interesting to see how many type-3 clones are reported in large code bases and how much value it would provide to the user to list them in the IDE. It would also be interesting to see how implementing a type-3 clone algorithm such as Baker's algorithm [7] would affect the speed of the incremental detection, and if any type-3 detection algorithms can be made more efficient for an incremental detection approach.

CCDetect-LSP identifies clones on a token level, unlike other tools which identifies clones on a line level [49]. It is more fine-grained to detect clones on a token level, and it could detect clones which would not be detected on a line level without needing to format the source code. However, it is possible that we could see a substantial performance and memory usage gain by instead fingerprinting entire lines instead of each token. This would drastically reduce the size of the fingerprint, which would make the suffix array updates faster and reduce the size of the dynamic extended suffix array and wavelet matrix. However, as there are a lot more unique lines in a code base than tokens, this would also drastically increase the alphabet size, σ, which will at least slow down operations on the wavelet matrix. This is a trade-off which would be interesting to explore further.

7.4 IDE and language agnostic clone detection

In this section we will briefly discuss how well CCDetect-LSP works in both an IDE and language agnostic setting.

IDE agnostic

CCDetect-LSP was implemented as a standalone program, which communicates with an IDE client via LSP. This means that the tool is IDE agnostic in the sense that any IDE which implements LSP can communicate with CCDetect-LSP and receive code clone information from it. With our implementation we are able to list clones in the form of diagnostics, and navigate between all the matches of a clone via code-actions and/or information embedded in the diagnostic. This demonstrates that LSP is capable of supporting the bare-minimum needed functionality of a clone detection tool.

In the previous chapter, we discussed that CCDetect-LSP is able to run in at least two IDEs which implement LSP, but we also saw that some LSP clients require more setup than others, and some LSP clients implement fewer features. This is the reality of working with a protocol which tries to unify multiple tools which do things differently. There will always be some IDEs which do not implement the full LSP protocol, but we believe that the features we have decided to use (diagnostics and code-actions) are good candidates to be implemented in most IDEs. Barros et al. [8] performed a study on LSP servers, which showed that out of 30 LSP servers, 28 of them implemented diagnostics, and 15 of them implemented code-actions. While this is not necessarily an indication of what the LSP clients implement, there is likely a correlation between what is generally offered by LSP servers, and what is therefore supported by LSP clients.

One downside to reporting code clones as diagnostics is that code clones are listed together with all other diagnostics reported by other LSP servers. If the project contains hundreds of clones

this could make the diagnostics list hard to use for any other usage than to list clones. A nice addition to the LSP protocol in this regard would be the possibility of categorizing diagnostics, and letting the client filter or show only diagnostics of a certain category. For CCDetect-LSP it would be nice to be able to have a separate view of only code clone diagnostics, which can be opened or closed as needed, and allow another LSP server to be the "main" source of diagnostics.

Another interesting feature which could be implemented with LSP is the concept of linked editing. Linked editing means being able to edit two or more locations at the same time, which could be used as an automatic way to apply an edit to all clones at once. LSP has since version 3.16 had such a feature, but we are not sure how fitting this feature is for code clone linked editing, as it seems targeted towards smaller refactoring operations such as renaming variables.

Language agnostic

The algorithm of CCDetect-LSP is also language agnostic. We have shown that as long as there exists a Tree-sitter grammar for a language,[2] it can be parsed and fingerprinted. When the source code has been fingerprinted, there is no longer any difference between detection for different languages. Therefore, CCDetect-LSP should work for any language, and in our testing, there is nothing that hinders CCDetect-LSP from working for any language we have tested, other than grammars not working correctly. We have tested CCDetect-LSP for multiple languages, and shown that there is very little configuration required to change the desired language. Therefore, we claim that CCDetect-LSP works very well as a language agnostic code analysis tool.

In the case of the C language, which did not work correctly because of a quirk of its grammar, it demonstrates a downside to going with the parser generator strategy. You cannot completely rely on every grammar to always work as you expect. It could be possible to forego parsing the source code, and only use a tokenizer to get the tokens. One could then use a language agnostic tokenization algorithm to avoid needing a grammar at all, and makes the algorithm truly language agnostic. However, this would not allow any sort of fragment selection, which likely leads to a lot of unnecessary code being analyzed for clones. This both slows down the detection process, but can also lead to a lot of noise in the clone list. For example in Java, a sorted list of imports at the beginning of the file could easily be cloned in multiple files, without any real possibility of being refactored to avoid this duplication. For this reason, we believe a "grammar-pluggable" approach with a parser generator such as Tree-sitter is a better approach than a completely language agnostic one, for the IDE scenario.

[2]Tree-sitter can parse most mainstream languages, see: `https://tree-sitter.github.io/tree-sitter/`

Chapter 8

Conclusion

This chapter will first list related and future work, then conclude the thesis.

8.1 Related work

This section will list related work which have explored similar topics, or work which could be useful in the previously listed future work.

Zibran et al. as previously cited, created the Eclipse plugin, called SimEclipse[42]. SimEclipse implements a suffix tree based algorithm to find code clones of a selected fragment in source code[49]. SimEclipse additionally implements many interesting features, such as visualization of clones in a tree-view, simultaneous editing of clones, and multiple features related to tracking the evolution of a clone.

Zhu et al. published a paper while this thesis was being written which shows a tool which implements grammar-pluggable clone detection [48]. This is the same concept which we have implemented in CCDetect-LSP where we achieve language agnostic clone detection by using a parser generator. They are able to identify all type-2 clones and has a 61% recall for type-3 clones. Their algorithm seems to be based on a completely different paradigm of clone detection than suffix trees and suffix arrays, which could be interesting to examine for its potential in an incremental setting. However, we should note that their evaluation shows that this type of clone detection is not very fast, so it is not clear if this approach is suitable in a "real-time" setting.

For our dynamic extended suffix array implementation, we built upon the ideas of Salson et al. [39, 30, 38] There have been other attempts to improve such an algorithm, such as Amir et al. [3]. They claim to have obtained a sub-linear time for queries in the suffix array, with a data structure where symbols can be substituted in the input string, in $O(\log^4(n))$ time. This would not be suitable for our clone detection algorithm, as we also require inserts and deletions. They also show a data structure with $O(\log^5(n))$ query-time for the inverted suffix array, which can maintain the inverted suffix array in $O(n^{\frac{2}{3}})$ time, and supports insertions and deletions as well. It would be interesting to see if this data structure could in any way improve the performance over our current approach, but this is questionable, as the data structure needs to support querying the suffix array, inverted suffix array and the LCP array efficiently, and preferably also support

inserting/deleting multiple characters at a time without performance problems.

8.2 Future work

In this section we list future work and research in order of what we consider most interesting to do more research on, to least.

Type-2 and type-3 clones

Finding type-2 clones is useful for CCDetect-LSP, but may not be of any interest for research, as applying an input normalization of certain token types is likely a satisfactory solution also for incremental detection.

Finding type-3 clones incrementally, could be more interesting to research. A classic algorithm to detect type-3 algorithms is to use sparse dynamic programming to find matches of type-1 and type-2 clones which are not exactly similar [7]. It would be interesting to look at such an algorithm in an incremental perspective, and determine if it would be beneficial to try and avoid computation of the same type-3 clones in each revision. Another approach could be to create a dynamic data structure which is simple to maintain between revisions, and allows extraction of type-3 clones in addition to type-1 and type-2 clones.

Optimal edit operations

A problem we discussed in chapter 5 is that our algorithm for computing edit operations is not optimal. The problem was to find an optimal amount of edit operations to turn one string into another, but we also preferred edit operations of more than one character, over many edit operations of one character. We implemented a simple algorithm which would aggregate edit operations of the same type which were consecutive in the edit matrix, but this could likely be optimized to output fewer operations with multiple characters in each. There is a trade-off where one could always reduce the problem to only two operations, delete the original string, and then insert the new string, but this might be more expensive than performing more edit operations, with a fewer number of characters. A possible approach is to assign a higher weight to starting a new edit operation than extending the previous operation, and build the edit matrix with these weights in mind, and using our algorithm to aggregate the edit operations afterwards with these weights. This could be difficult if Hirschberg's algorithm is still used, but the regular Wagner-Fischer algorithm could be used in a majority of cases if the memory usage of the regular Wagner-Fischer matrix is not too big.

Refactoring of clones

Another interesting topic is the refactoring of clones. Automatic refactoring of clones has not been thoroughly researched, but automatic refactoring of code clones could be beneficial to practically all software. Some literature exists on the topic [5, 20], especially targeted at object-oriented programming languages. For CCDetect-LSP it would be interesting to look at either combining the clone detection with existing refactoring tools from other LSP clients, or to implement refactoring code-actions in CCDetect-LSP itself, which automatically refactors to remove the selected clone. However, this would be a hard to achieve while CCDetect-LSP is still a language agnostic tool, as different languages have completely different paradigms to perform refactoring on.

Compressing data structures

A problem with CCDetect-LSP in its current state, is its memory usage. As mentioned, when the size of the code base grows, the memory usage of CCDetect-LSP grows as well, and might reach a point where it is not feasible for lower-end computers to run it simultaneously with other applications, which is often the case in an IDE scenario. As mentioned, it is possible that reducing the size of the fingerprint by fingerprinting entire lines instead of tokens could improve the memory usage, but another potential solution is to compress the memory consuming data structures.

The wavelet matrix can be compressed to zero-order entropy, but Claude et al. explain that this compressed wavelet matrix is likely not as efficient as a compressed wavelet tree [12]. It might be worthwhile to implement either a compressed wavelet matrix or a compressed huffman-shaped wavelet tree, and test how this affects both memory usage and performance. It is likely that the performance will be slower with this compressed data structure, but the memory usage will likely improve as well.

It is also possible to compress the dynamic extended suffix array data structure as explained by Salson et al. [39]. This is done by sampling (storing) only some values of the dynamic permutation at a time, and computing the values at unsampled positions by using the sampled values. This would again slow down the performance of accessing the data structure, as computing the values at unsampled positions is slower, but it will also reduce the number of values stored, and therefore the memory usage will be better.

Lazy LCP array updates

In the suffix array update phase, for each edit operation we computed in the previous phase, we run a suffix array update. In each one of these updates, we determine which positions in the LCP array need to be updated, and then compute the LCP value at that position. It is possible that we can avoid some work here if multiple of these edit operations lead to updating the same LCP positions, as we can delay the computation of those LCP values until after all the edit operations have updated the suffix array. This would be of interest for algorithms which utilizes the dynamic extended suffix array algorithm in general, if multiple edits are applied.

8.3 Conclusion

The results demonstrate that the incremental algorithm of CCDetect-LSP scales better than the other two algorithms in terms of time, at least in the IDE scenario where edits are relatively small. The incremental algorithm was able to process large edits for code bases of over 5MLOC in under 2 seconds, and faster for smaller edits. We have demonstrated that our novel application of dynamic extended suffix arrays is a suitable and fast approach to clone detection in the IDE scenario, and can give comparable and better results than tools which implement a dynamic suffix tree. It also outperforms other incremental detection algorithms in terms of memory. Though memory consumption is the main limitation for practical use, CCDetect-LSP still outperforms other comparable tools in the IDE scenario.

We have demonstrated that LSP can implement the bare-minimum features needed for a clone detection tool, but some extensions would be useful in order to separate clone detection from other types of diagnostics.

Finally, we have demonstrated that language agnostic clone detection using a parser generator works and allows sufficient selection of fragments for flexible analysis, and to avoid listing unwanted clones. We have also discussed that a language agnostic approach is more challenging in terms of normalizing inputs and therefore detection of type-2 clones.

Bibliography

[1] Mohamed Ibrahim Abouelhoda, Stefan Kurtz, and Enno Ohlebusch. "Replacing suffix trees with enhanced suffix arrays". In: *Journal of Discrete Algorithms* 2.1 (2004). The 9th International Symposium on String Processing and Information Retrieval, pp. 53–86. ISSN: 1570-8667. DOI: https://doi.org/10.1016/S1570-8667(03)00065-0. URL: https://www.sciencedirect.com/science/article/pii/S1570866703000650.

[2] Raihan Al-Ekram et al. "Cloning by accident: an empirical study of source code cloning across software systems". In: *2005 International Symposium on Empirical Software Engineering (ISESE 2005), 17-18 November 2005, Noosa Heads, Australia*. IEEE Computer Society, 2005, pp. 376–385. DOI: 10.1109/ISESE.2005.1541846. URL: https://doi.org/10.1109/ISESE.2005.1541846.

[3] Amihood Amir and Itai Boneh. *Dynamic Suffix Array with Sub-linear update time and Poly-logarithmic Lookup Time*. 2021. arXiv: 2112.12678. URL: https://arxiv.org/abs/2112.12678.

[4] Paris Avgeriou et al. "Managing Technical Debt in Software Engineering (Dagstuhl Seminar 16162)". In: *Dagstuhl Reports* 6.4 (2016), pp. 110–138. DOI: 10.4230/DagRep.6.4.110. URL: https://doi.org/10.4230/DagRep.6.4.110.

[5] Simon Baars and Ana Oprescu. *Towards Automated Refactoring of Code Clones in Object-Oriented Programming Languages*. EasyChair Preprint no. 1278. EasyChair, 2019.

[6] Brenda S. Baker. "A program for identifying duplicated code". In: *Computer Science and Statistics: Proc. Symp. on the Interface*. Mar. 1992, pp. 49–57. URL: http://citeseer.nj.nec.com/baker92program.html.

[7] Brenda S. Baker and Raffaele Giancarlo. "Sparse Dynamic Programming for Longest Common Subsequence from Fragments". In: *Journal of Algorithms* 42.2 (2002), pp. 231–254. ISSN: 0196-6774. DOI: https://doi.org/10.1006/jagm.2002.1214. URL: https://www.sciencedirect.com/science/article/pii/S0196677402912149.

[8] Djonathan Barros et al. "Editing Support for Software Languages: Implementation Practices in Language Server Protocols". In: *Proceedings of the 25th International Conference on Model Driven Engineering Languages and Systems*. MODELS '22. Montreal, Quebec, Canada: Association for Computing Machinery, 2022, pp. 232–243. ISBN: 9781450394666. DOI: 10.1145/3550355.3552452. URL: https://doi.org/10.1145/3550355.3552452.

[9] Ira Baxter et al. "Clone Detection Using Abstract Syntax Trees." In: vol. 368-377. Jan. 1998, pp. 368–377. DOI: 10.1109/ICSM.1998.738528.

[10] Max Brunsfeld. *Tree-sitter*. https://tree-sitter.github.io/tree-sitter/. Accessed: 2022-04-16.

[11] Michael Burrows and David J. Wheeler. “A Block-sorting Lossless Data Compression Algorithm”. In: 1994.

[12] Francisco Claude, Gonzalo Navarro, and Alberto Ordóñez Pereira. “The wavelet matrix: An efficient wavelet tree for large alphabets”. In: *Inf. Syst.* 47 (2015), pp. 15–32. DOI: 10.1016/j.is.2014.06.002. URL: https://doi.org/10.1016/j.is.2014.06.002.

[13] Thomas H. Cormen et al. *Introduction to Algorithms, Third Edition*. 3rd. The MIT Press, 2009. ISBN: 0262033844.

[14] P.B. Crosby. *Quality is Free: The Art of Making Quality Certain*. New American Library, 1980. ISBN: 9780451624680. URL: https://books.google.no/books?id=3TMQt73LDooC.

[15] Martin Fowler. *Refactoring - Improving the Design of Existing Code*. Addison Wesley object technology series. Addison-Wesley, 1999. ISBN: 978-0-201-48567-7. URL: http://martinfowler.com/books/refactoring.html.

[16] Erich Gamma et al. *Design Patterns: Elements of Reusable Object-Oriented Software*. 1st ed. Addison-Wesley Professional, 1994. ISBN: 0201633612. URL: http://www.amazon.com/Design-Patterns-Elements-Reusable-Object-Oriented/dp/0201633612/ref=ntt_at_ep_dpi_1.

[17] Nils Göde and Rainer Koschke. “Incremental Clone Detection”. In: *2009 13th European Conference on Software Maintenance and Reengineering*. 2009, pp. 219–228. DOI: 10.1109/CSMR.2009.20.

[18] Roberto Grossi, Ankur Gupta, and Jeffrey Scott Vitter. “High-order entropy-compressed text indexes”. In: *Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms, January 12-14, 2003, Baltimore, Maryland, USA*. ACM/SIAM, 2003, pp. 841–850. URL: http://dl.acm.org/citation.cfm?id=644108.644250.

[19] Dan Gusfield. “Algorithms on stings, trees, and sequences: Computer science and computational biology”. In: *Acm Sigact News* 28.4 (1997), pp. 41–60.

[20] Yoshiki Higo. “Identifying Refactoring-Oriented Clones and Inferring How They Can Be Merged”. In: *Code Clone Analysis: Research, Tools, and Practices*. Ed. by Katsuro Inoue and Chanchal K. Roy. Singapore: Springer Singapore, 2021, pp. 183–196. ISBN: 978-981-16-1927-4. DOI: 10.1007/978-981-16-1927-4_13. URL: https://doi.org/10.1007/978-981-16-1927-4_13.

[21] D. S. Hirschberg. “A Linear Space Algorithm for Computing Maximal Common Subsequences”. In: *Commun. ACM* 18.6 (June 1975), pp. 341–343. ISSN: 0001-0782. DOI: 10.1145/360825.360861. URL: https://doi.org/10.1145/360825.360861.

[22] Benjamin Hummel et al. “Index-based code clone detection: incremental, distributed, scalable”. In: *2010 IEEE International Conference on Software Maintenance*. 2010, pp. 1–9. DOI: 10.1109/ICSM.2010.5609665.

[23] Katsuro Inoue. “Introduction to Code Clone Analysis”. In: *Code Clone Analysis: Research, Tools, and Practices*. Ed. by Katsuro Inoue and Chanchal K. Roy. Singapore: Springer Singapore, 2021, pp. 3–27. ISBN: 978-981-16-1927-4. DOI: 10.1007/978-981-16-1927-4_1. URL: https://doi.org/10.1007/978-981-16-1927-4_1.

[24] G. Jacobson. “Space-efficient static trees and graphs”. In: *30th Annual Symposium on Foundations of Computer Science*. 1989, pp. 549–554. DOI: 10.1109/SFCS.1989.63533.

[25] Lingxiao Jiang et al. “DECKARD: Scalable and Accurate Tree-Based Detection of Code Clones”. In: *29th International Conference on Software Engineering (ICSE'07)*. 2007, pp. 96–105. DOI: 10.1109/ICSE.2007.30.

[26] Stephen H. Kan. *Metrics and Models in Software Quality Engineering*. 2nd. USA: Addison-Wesley Longman Publishing Co., Inc., 2002. ISBN: 0201729156.

[27] Juha Kärkkäinen. "Suffix Array Construction". In: *Encyclopedia of Algorithms*. Ed. by Ming-Yang Kao. New York, NY: Springer New York, 2016, pp. 2141–2144. ISBN: 978-1-4939-2864-4. DOI: 10.1007/978-1-4939-2864-4_412. URL: https://doi.org/10.1007/978-1-4939-2864-4_412.

[28] Shinji Kawaguchi et al. "SHINOBI: A Tool for Automatic Code Clone Detection in the IDE". In: *16th Working Conference on Reverse Engineering, WCRE 2009, 13-16 October 2009, Lille, France*. Ed. by Andy Zaidman, Giuliano Antoniol, and Stéphane Ducasse. IEEE Computer Society, 2009, pp. 313–314. DOI: 10.1109/WCRE.2009.36. URL: https://doi.org/10.1109/WCRE.2009.36.

[29] Bernard Lang. "Deterministic Techniques for Efficient Non-Deterministic Parsers". In: *Automata, Languages and Programming*. Ed. by Jacques Loeckx. Berlin, Heidelberg: Springer Berlin Heidelberg, 1974, pp. 255–269. ISBN: 978-3-662-21545-6.

[30] M. Léonard, L. Mouchard, and M. Salson. "On the number of elements to reorder when updating a suffix array". In: *Journal of Discrete Algorithms* 11 (2012). Special issue on Stringology, Bioinformatics and Algorithms, pp. 87–99. ISSN: 1570-8667. DOI: https://doi.org/10.1016/j.jda.2011.01.002. URL: https://www.sciencedirect.com/science/article/pii/S1570866711000037.

[31] Veli Mäkinen and Gonzalo Navarro. "Rank and select revisited and extended". In: *Theoretical Computer Science* 387.3 (2007). The Burrows-Wheeler Transform, pp. 332–347. ISSN: 0304-3975. DOI: https://doi.org/10.1016/j.tcs.2007.07.013. URL: https://www.sciencedirect.com/science/article/pii/S0304397507005300.

[32] Microsoft. *Language Server Protocol*. https://microsoft.github.io/language-server-protocol/. Accessed: 2023-02-17.

[33] Tung Thanh Nguyen et al. "Scalable and incremental clone detection for evolving software". In: *25th IEEE International Conference on Software Maintenance (ICSM 2009), September 20-26, 2009, Edmonton, Alberta, Canada*. IEEE Computer Society, 2009, pp. 491–494. DOI: 10.1109/ICSM.2009.5306283. URL: https://doi.org/10.1109/ICSM.2009.5306283.

[34] Ge Nong, Sen Zhang, and Wai Hong Chan. "Two Efficient Algorithms for Linear Time Suffix Array Construction". In: *IEEE Trans. Computers* 60.10 (2011), pp. 1471–1484. DOI: 10.1109/TC.2010.188. URL: https://doi.org/10.1109/TC.2010.188.

[35] Chaiyong Ragkhitwetsagul and Jens Krinke. "Siamese: scalable and incremental code clone search via multiple code representations". In: *Empir. Softw. Eng.* 24.4 (2019), pp. 2236–2284. DOI: 10.1007/s10664-019-09697-7. URL: https://doi.org/10.1007/s10664-019-09697-7.

[36] M. Rieger, S. Ducasse, and M. Lanza. "Insights into system-wide code duplication". In: *11th Working Conference on Reverse Engineering*. 2004, pp. 100–109. DOI: 10.1109/WCRE.2004.25.

[37] Chanchal Kumar Roy, James R. Cordy, and Rainer Koschke. "Comparison and evaluation of code clone detection techniques and tools: A qualitative approach". In: *Sci. Comput. Program.* 74.7 (2009), pp. 470–495. DOI: 10.1016/j.scico.2009.02.007. URL: https://doi.org/10.1016/j.scico.2009.02.007.

[38] M. Salson et al. “A four-stage algorithm for updating a Burrows–Wheeler transform”. In: *Theoretical Computer Science* 410.43 (2009). String Algorithmics: Dedicated to Professor Maxime Crochemore on the occasion of his 60th birthday, pp. 4350–4359. ISSN: 0304-3975. DOI: https://doi.org/10.1016/j.tcs.2009.07.016. URL: https://www.sciencedirect.com/science/article/pii/S0304397509004770.

[39] M. Salson et al. “Dynamic extended suffix arrays”. In: *Journal of Discrete Algorithms* 8.2 (2010). Selected papers from the 3rd Algorithms and Complexity in Durham Workshop ACiD 2007, pp. 241–257. ISSN: 1570-8667. DOI: https://doi.org/10.1016/j.jda.2009.02.007. URL: https://www.sciencedirect.com/science/article/pii/S1570866709000343.

[40] Jeffrey Svajlenko and Chanchal K. Roy. “BigCloneBench”. In: *Code Clone Analysis*. Ed. by Katsuro Inoue and Chanchal K. Roy. Springer Singapore, 2021, pp. 93–105. DOI: 10.1007/978-981-16-1927-4_7. URL: https://doi.org/10.1007/978-981-16-1927-4%5C_7.

[41] Jeffrey Svajlenko and Chanchal K. Roy. “BigCloneEval: A Clone Detection Tool Evaluation Framework with BigCloneBench”. In: *2016 IEEE International Conference on Software Maintenance and Evolution (ICSME)*. 2016, pp. 596–600. DOI: 10.1109/ICSME.2016.62.

[42] Md Sharif Uddin, Chanchal K. Roy, and Kevin A. Schneider. “Towards Convenient Management of Software Clone Codes in Practice: An Integrated Approach”. In: *Proceedings of the 25th Annual International Conference on Computer Science and Software Engineering*. CASCON ’15. Markham, Canada: IBM Corp., 2015, pp. 211–220.

[43] Esko Ukkonen. “On approximate string matching”. In: *Foundations of Computation Theory*. Ed. by Marek Karpinski. Berlin, Heidelberg: Springer Berlin Heidelberg, 1983, pp. 487–495. ISBN: 978-3-540-38682-7.

[44] Esko Ukkonen. “On-line construction of suffix trees”. In: *Algorithmica* 14.3 (1995), pp. 249–260.

[45] Robert A. Wagner and Michael J. Fischer. “The String-to-String Correction Problem”. In: *J. ACM* 21.1 (Jan. 1974), pp. 168–173. ISSN: 0004-5411. DOI: 10.1145/321796.321811. URL: https://doi.org/10.1145/321796.321811.

[46] Tim A. Wagner. “Practical Algorithms for Incremental Software Development Environments”. PhD thesis. EECS Department, University of California, Berkeley, Mar. 1998. URL: http://www2.eecs.berkeley.edu/Pubs/TechRpts/1998/5885.html.

[47] Zhipeng Xue et al. “SEED: Semantic Graph based Deep detection for type-4 clone”. In: *CoRR* abs/2109.12079 (2021). arXiv: 2109.12079. URL: https://arxiv.org/abs/2109.12079.

[48] Wenqing Zhu et al. “MSCCD: Grammar Pluggable Clone Detection Based on ANTLR Parser Generation”. In: *Proceedings of the 30th IEEE/ACM International Conference on Program Comprehension*. ACM, May 2022. DOI: 10.1145/3524610.3529161. URL: https://doi.org/10.1145%2F3524610.3529161.

[49] Minhaz F. Zibran and Chanchal K. Roy. “IDE-Based Real-Time Focused Search for near-Miss Clones”. In: *Proceedings of the 27th Annual ACM Symposium on Applied Computing*. SAC ’12. Trento, Italy: Association for Computing Machinery, 2012, pp. 1235–1242. ISBN: 9781450308571. DOI: 10.1145/2245276.2231970. URL: https://doi.org/10.1145/2245276.2231970.

Chapter 9

Appendix

The appendix contains supplementary figures, logs and algorithms which do not naturally fit into the thesis. Algorithms in this chapter are generally very verbose, and it is recommended to read them alongside their respective explanations in the implementation chapters.

```
./evaluateTool --min-tokens=100 -s=100 --output=ccdetect.report -t=1

-- Versioning --
 BigCloneEval: Release Version 0.1
BigCloneBench: Version 1.0 (2016-06-19)
-- Selected Clones --
         Min Lines: null
         Max Lines: null
        Min Tokens: 100
        Max Tokens: null
  Min Pretty Lines: null
  Max Pretty Lines: null
        Min Judges: null
    Min Confidence: null
          Sim Type: BOTH
Minimum Similarity:

-- Clone Matcher --
Coverage Matcher.  Coverage Ratio = 0.7, minimum raito is of reference clone: null

================================================================================
All Functionalities
================================================================================
-- Recall Per Clone Type (type: numDetected / numClones = recall) --
              Type-1: 23205 / 23210 = 0.9997845756139595
              Type-2: 55 / 3547 = 0.015506061460389062
      Type-2 (blind): 2 / 245 = 0.00816326530612245
 Type-2 (consistent): 53 / 3302 = 0.016050878255602665

-- Inter-Project Recall Per Clone Type (type: numDetected / numClones = recall)  --
              Type-1: 4620 / 4625 = 0.9989189189189189
              Type-2: 24 / 3185 = 0.0075353218210361065
      Type-2 (blind): 0 / 223 = 0.0
 Type-2 (consistent): 24 / 2962 = 0.008102633355840648

-- Intra-Project Recall Per Clone Type (type: numDetected / numClones = recall) --
-- Recall Per Clone Type --
              Type-1: 18585 / 18585 = 1.0
              Type-2: 31 / 362 = 0.0856353591160221
      Type-2 (blind): 2 / 22 = 0.09090909090909091
 Type-2 (consistent): 29 / 340 = 0.08529411764705883
```

Figure 9.1: BigCloneEval evaluation report on CCDetect-LSP

```
Algorithm WagnerFischerOperations(S1, S2, matrix)
    operations ← Empty list
    currentOperation ← None operation
    lastOperationIndex ← −1
    x ← Len(matrix) − 1
    y ← Len(matrix[x]) − 1

    while x > 0 and y > 0 do
        subValue ← matrix[x − 1][y − 1]
        delValue ← matrix[x − 1][y]
        insValue ← matrix[x][y − 1]
        if subValue ≤ delValue and subValue ≤ insValue and  then
            if subValue ≠ matrix[x][y] then
                if currentOperation is not a substitute or lastOperationIndex ≠ x then
                    currentOperation ← New substitute operation with position x − 1
                    Insert(operations, 0, currentOperation) // Add operation at index 0
                end
                else
                    currentOperation.position ← currentOperation.position − 1
                end
            end
            Add S2[y − 1] to beginning of currentOperation.chars
            lastOperationIndex ← x − 1
            x ← x − 1
            y ← y − 1
        end
        else if insValue ≤ delValue then
            if currentOperation is not an insert or lastOperationIndex ≠ x then
                currentOperation ← New delete operation with position y − 1
                Insert(operations, 0, currentOperation) // Add operation at index 0
            end
            Add S1[x − 1] to beginning of currentOperation.chars
            lastOperationIndex ← x − 1
            y ← y − 1
        end
        else
            if currentOperation is not a delete or lastOperationIndex ≠ x then
                currentOperation ← New insert operation with position x − 1
                Insert(operations, 0, currentOperation) // Add operation at index 0
            end
            Add S1[x − 1] to beginning of currentOperation.chars
            lastOperationIndex ← x − 1
            x ← x − 1
        end
    end
    return operations
```

Algorithm 18: Compute aggregated edit operations from Wagner-Fischer matrix

```
Algorithm WagnerFischerEditDistance(S_1, S_2)
    n ← Len(S_1)
    m ← Len(S_2)
    matrix ← new array[n + 1][m + 1]

    for i from 0 to n do
        matrix[i][0] = i
    end

    for i from 0 to m do
        matrix[0][i] = i
    end

    for i from 1 to n do
        for j from 1 to m do
            if S_1[i − 1] = S_2[j − 1] then
                matrix[i][j] = matrix[i − 1][j − 1]
            end
            else
                delete ← matrix[i − 1][j]
                insert ← matrix[i][j − 1]
                substitute ← matrix[i − 1][j − 1]

                matrix[i][j] = Min(insert, delete, substitute) + 1
            end
        end
    end
    return matrix
```

Algorithm 19: Fill edit distance matrix using Wagner-Fischer algorithm