

UNIVERSIDAD MAYOR DE SAN ANDRÉS
FACULTAD DE CIENCIAS PURAS Y NATURALES
CARRERA DE INFORMATICA



TESIS DE GRADO
DISEÑO DE UN ALGORITMO PARA LA DETECCIÓN DE
SIMILITUD ENTRE CÓDIGOS FUENTE

Para optar por el Título de Licenciatura en Informática

Mención: Ciencias de la Computación

Postulante: Univ. Edson Eddy Lecoña Zarate

Tutor: M.Sc. Jorge Humberto Terán Pomier

La Paz - Bolivia

2022

Índice general

| | |
|---|----------|
| 1. MARCO REFERENCIAL | 4 |
| 1.1. Introducción | 4 |
| 1.2. Problema | 5 |
| 1.2.1. Antecedentes del problema | 5 |
| 1.2.2. Planteamiento del problema | 7 |
| 1.2.3. Formulación del problema | 7 |
| 1.3. Objetivos | 7 |
| 1.3.1. Objetivo general | 7 |
| 1.3.2. Objetivos específicos | 7 |
| 1.4. Hipótesis | 8 |
| 1.4.1. Variables Independientes | 8 |
| 1.4.2. Variables Dependientes | 8 |
| 1.5. Justificaciones | 8 |
| 1.5.1. Justificación Social | 8 |
| 1.5.2. Justificación Económica | 8 |
| 1.5.3. Justificación Tecnológica | 9 |
| 1.5.4. Justificación Científica | 9 |
| 1.6. Alcances y Limites | 9 |
| 1.6.1. Alcance Sustancial | 9 |
| 1.6.2. Alcance Espacial | 9 |
| 1.6.3. Alcance Temporal | 10 |
| 1.7. Metodología | 10 |
| 1.7.1. Metodología Experimental | 10 |

| | |
|---|-----------|
| 2. MARCO TEÓRICO | 11 |
| 2.1. Conceptos de lenguajes de programación | 11 |
| 2.1.1. Lenguaje de programación | 11 |
| 2.2. Conceptos de código fuente | 11 |
| 2.2.1. Código fuente | 11 |
| 2.2.2. Ofuscación de código fuente | 12 |
| 2.2.3. Métodos de ofuscación de código fuente | 12 |
| 2.3. Conceptos de compiladores | 13 |
| 2.3.1. Procesador de lenguaje | 13 |
| 2.3.2. Análisis Léxico | 14 |
| 2.3.3. Análisis Sintáctico | 15 |
| 2.3.4. Análisis Semántico | 16 |
| 2.3.5. Árboles de sintaxis abstracta | 16 |
| 2.4. Conceptos de algorítmica | 18 |
| 2.4.1. Programación dinámica | 18 |
| 2.4.2. Distancia de edición de árbol | 19 |
| 2.4.3. Conjuntos Disjuntos | 19 |
| 2.5. Herramientas para la detección de similitud de código | 20 |
| 2.5.1. Algoritmos utilizados para la detección de similitud | 20 |
| 2.5.2. Detección léxica | 20 |
| Bibliografía | 21 |

Índice de figuras

| | |
|---|----|
| 2.1. Un compilador | 13 |
| 2.2. Fases de un compilador | 14 |
| 2.3. Traducción de una asignación | 17 |
| 2.4. Árbol | 18 |

CAPÍTULO 1: MARCO REFERENCIAL

1.1. Introducción

[Cheers *et al.*, 2021] Explica que la identificación de similitud entre códigos fuente puede servir para varios propósitos, entre ellos están el estudio de la evolución de código fuente de un proyecto, detección de prácticas de plagio, detección de prácticas de reutilización, extracción de código para “refactorización” del mismo y seguimiento de defectos para su corrección.

Los estudiantes durante su proceso de formación elaboran trabajos, proyectos, tareas y ejercicios de programación escritos en un lenguaje de programación, estas actividades se realiza de forma individual o grupal, cuando estas actividades se desarrollan de forma individual los estudiantes deben ser conscientes de que todo trabajo entregado debe ser de su creación, pues tiene como función medir la capacidad de resolución de problemas y el enfoque lógico y otros. Por ello encontrar similitud en trabajos de programación presentados por los estudiantes, puede ser identificado como plagio.

Una de las formas para detectar el plagio en los trabajos de programación consiste en realizar la comparación entre los trabajos entregados por los estudiantes de la materia. De a modo de obtener una lista de estudiantes que tienen trabajos similares. Realizar la comparación de los trabajos manualmente puede llegar a ser un trabajo moroso, por lo cual contar con una herramienta de software que realice las comparaciones de trabajos de forma automática es de gran utilidad.

En la actualidad existen diferentes herramientas de software que aplican métodos para detección de similitud entre códigos fuente, a partir de las características se pudo evidenciar que estas presentan deficiencias como ser la obsolescencia, sistemas cerrados (sin código abierto), un proceso complejo de evaluación de similitud, la incapacidad para utilizar o no una gran base de información. Esto se debe a que fueron diseñadas para detectar plagio

entre un grupo pequeño de archivos de código fuente.

Por lo cual contar con un sistema para la detección de similitud entre códigos fuente llega a ser útil para detectar plagio en trabajos de cátedra presentados por estudiantes. El presente trabajo de tesis se centra en el diseño e implementación de un algoritmo para la detección de similitud entre códigos fuente que tenga un buen desempeño, en términos de tiempo de ejecución, espacio de memoria ocupado y precisión.

1.2. Problema

1.2.1. Antecedentes del problema

Se analizaron herramientas para la detección de similitud entre código fuente más populares, a continuación se dará breve explicación de las características que tienen estas.

Sherlock

Es una herramienta de código abierto, que trabaja con código escrito en los lenguajes Java, C y texto natural. Esta herramienta no cuenta con una interfaz gráfica. Fue desarrollada por la Universidad de Sydney. Los resultados arrojados se basan en un porcentaje que se corresponde con las similitudes encontradas. El porcentaje 0 % significa que no hay similitudes, y 100 % significa que hay muchas posibilidades de que tengan partes iguales. Al no utilizar el documento en su totalidad, no se puede afirmar que sean completamente iguales. Sólo trabaja con archivos locales, no busca similitudes en Internet [Díaz *et al.*, 2007].

Simian

Esta herramienta no es de código abierto. Identifica la duplicación de códigos escritos en JAVA, C, C++, COBOL, Ruby, JSP, ASP, HTML, XML, Visual Basic y texto natural. Fue desarrollada por una consultora de Australia llamada REDHILL. No cuenta con una interfaz gráfica, por lo tanto, el ingreso de los parámetros es a través de línea de comando. Sólo trabaja con archivos locales, no busca similitudes en Internet [Díaz *et al.*, 2007].

Jplag

Esta herramienta no es de código abierto, sólo permite la programación del cliente, pero el servidor es el que realiza las operaciones de comparación; se encuentra en la Universidad de Karlsruhe de Alemania, en la cual fue desarrollada. Trabaja con los siguientes lenguajes C, C++, Java, Scheme y texto natural. Fue desarrollada por la Universidad de Karlsruhe de Alemania. Analiza la estructura y sintaxis del código, no comparando texto. Este sistema sólo trabaja con archivos locales, no busca similitudes en Internet. Su arquitectura de trabajo es Cliente/Servidor. La interfaz es a través de línea de comando, o, puede implementarse un cliente propio. Para el manejo de los archivos enviados, puede utilizarse el cliente brindado por la universidad que lo desarrolló o bien crear un cliente propio, como mencionamos anteriormente [Díaz *et al.*, 2007].

Tester SIM

Es una herramienta de código abierto. Trabaja con los lenguajes C, JAVA, Lisp, Modula 2, Pascal y texto natural. Está desarrollada por la Universidad de Amsterdam. No cuenta con una interfaz gráfica, se ingresan los distintos parámetros por línea de comando. Los resultados brindados por la misma aparecen separados en dos columnas, mostrando en cada una las porciones de código iguales. Sólo trabaja con búsquedas de similitudes en archivos locales. También se puede procesar un conjunto de archivos viejos contra un conjunto de archivos nuevos [Díaz *et al.*, 2007].

MOSS

Software creado por el profesor Alex Aiken en la universidad de Stanford, es el primer servicio que inició en la web, siendo una referencia a nivel mundial. Este, permite comparar hasta 250 archivos en 25 lenguajes de programación [Hage y Rademaker, 2010]. Aunque no tiene licencia como software libre, su uso dentro del ambiente académico es gratuito y ofrecido desde un servidor de Stanford. Es de difícil su configuración y tiene poca documentación [Pachón, 2019].

1.2.2. Planteamiento del problema

Los estudiantes de programación durante su proceso de formación elaboran trabajos, proyectos, tareas, ejercicios, etc. de programación, escritos en algún lenguaje de programación. Estas actividades deben realizarse de forma individual dado que sirven para medir la capacidad de resolución de problemas y el enfoque lógico de los estudiantes. Por ello encontrar similitud en trabajos de programación presentados por estudiantes puede ser identificado como plagio de código fuente.

La forma mas simple para detectar plagio en trabajos de programación, consiste en realizar la comparación de forma manual entre todos los trabajos presentados, de modo de obtener una lista de estudiantes que tienen trabajos similares. Realizar esta comparación llega a ser una trabajo moroso, por lo cual contar con una herramienta de software que realice las comparaciones de forma automática y genere la una lista de trabajos similares es de gran utilidad.

En la actualidad existen herramientas de software para detectar similitud entre códigos fuente. Según sus características se observo que estas presentan deficiencias como la obsolescencia, sin código abierto, incapacidad de utilizar gran base de información.

1.2.3. Formulación del problema

¿El algoritmo JTEL tiene mejor desempeño frente a otras herramientas en la detección de similitud de códigos fuente en trabajos de cátedra?

1.3. Objetivos

1.3.1. Objetivo general

Diseñar e implementar el algoritmo JTEL para la detección de similitud entre códigos fuente.

1.3.2. Objetivos específicos

- Estudiar los métodos para la detección de similitud entre códigos fuente existentes.

- Estudiar los algoritmos utilizados en los métodos para la detección de similitud de código fuente.
- Redactar las especificaciones para el algoritmo JTEL.
- Evaluar el desempeño del algoritmo JTEL realizando las pruebas en trabajos de programación presentado por estudiantes.

1.4. Hipótesis

El algoritmo JTEL implementado para la detección de similitud entre códigos fuente obtiene mejores resultados frente a otras herramientas para la detección de similitud.

1.4.1. Variables Independientes

- El algoritmo JTEL para la detección de similitud entre códigos fuente.

1.4.2. Variables Dependientes

- Resultados más precisos respecto a la detección de similitud entre códigos fuente de trabajos de cátedra frente a otras herramientas.

1.5. Justificaciones

1.5.1. Justificación Social

El algoritmo JTEL ahorrará el tiempo de docentes de instituciones académicas en la detección de plagio en trabajos de programación presentados por estudiantes, evitando que los docentes realicen la comparación de trabajos de programación de forma manual.

1.5.2. Justificación Económica

El algoritmo JTEL para la detección de similitud entre códigos fuente sera de código abierto, por lo cual permitirá que se desarrollen otros software a bajo costo.

1.5.3. Justificación Tecnológica

El algoritmo JTEL para la detección de similitud entre códigos fuente, se puede implementar en jueces de programación para identificar los envíos similares de los usuarios del juez.

1.5.4. Justificación Científica

Con la utilización de trabajos de programación presentados por estudiantes se medirá el desempeño de los métodos existentes para la detección de similitud entre códigos fuente. Con los resultados obtenidos se determinará cuál es el método más eficiente en términos de tiempo de ejecución, espacio de memoria ocupado y precisión.

1.6. Alcances y Limites

1.6.1. Alcance Sustancial

- Se diseñará e implementará en Python un algoritmo JTEL para la detección de similitud entre códigos fuente.
- Se realizaran las pruebas de trabajos de cátedra de código fuente en Python.
- Se dejará de lado la comprobación de correctitud de los códigos fuente.
- Se realizarán pruebas para medir el tiempo, espacio de memoria ocupado y eficiencia del algoritmo.
- Se dejará de lado el estudio de métodos para la detección de similitud que aplican técnicas de Inteligencia Artificial.

1.6.2. Alcance Espacial

- Se realizaran las pruebas del algoritmo en una computadora intel i3 de Décima generación con 8GB de RAM y 512GB de Disco Duro.

1.6.3. Alcance Temporal

- El tiempo de ejecución que el algoritmo estará limitado a un minuto. Por cada conjunto de trabajos evaluados.

1.7. Metodología

1.7.1. Metodología Experimental

Para el desarrollo del trabajo de investigación se utilizara la metodología científica experimental, esta investigación nos permite la manipulación de una o mas variables. De modo que al seguir las siguientes etapas ayudara a cumplir con los objetivos propuestos.

1. **Recopilación de la información.** En esta etapa se recopilara información necesaria y estudiara los temas.
2. **Diseño del algoritmo.** En esta etapa se diseñara el algoritmo tomando en cuenta los alcances y limites.
3. **Pruebas de funcionamiento del algoritmo en trabajos de cátedra.** En esta etapa se Realizara pruebas en trabajos de cátedra presentados por estudiantes.
4. **Análisis de los resultados obtenidos.** En esta etapa se analizaran los datos y se los comparara frente a otras herramientas.
5. **Conclusiones.** En esta etapa se realizara las conclusiones, se presentaran los resultados finales, y recomendaciones respecto a la investigación.

CAPÍTULO 2: MARCO TEÓRICO

2.1. Conceptos de lenguajes de programación

2.1.1. Lenguaje de programación

Los lenguajes de programación son notaciones que describen los cálculos a las personas y las máquinas. Nuestra percepción del mundo en que vivimos depende de los lenguajes de programación, ya que todo el software que se ejecuta en todas las computadoras se escribió en algún lenguaje de programación [Aho, 2008].

Un lenguaje de programación es un lenguaje formal, que mediante un conjunto de instrucciones permite a un programador crear programas. El lenguaje de programación es un sistema estructurado de comunicación conformado por conjuntos de símbolos, palabras clave, reglas semánticas y sintácticas, las cuales sirven para el entendimiento entre un programador y una máquina.

2.2. Conceptos de código fuente

2.2.1. Código fuente

En informática, se denomina código fuente a la conjunto de líneas de texto que escritas por un programador. Estas líneas de texto representan instrucciones en un lenguaje de programación. Las instrucciones representan los pasos que debe seguir la computadora para la ejecución de un programa específico.

El código fuente no es directamente ejecutable por la computadora, este debe ser traducido a otro lenguaje de modo que la computadora pueda interpretarlo. En la traducción se usan compiladores, ensambladores, interpretes y otros.

2.2.2. Ofuscación de código fuente

En computación, la ofuscación se refiere al acto deliberado de realizar un cambio no destructivo, ya sea en el código fuente de un programa informático, en el código intermedio o en el código máquina cuando el programa está en forma compilada o binaria. Es decir, se cambia el código se enrevesa manteniendo el funcionamiento original, para dificultar su entendimiento. De esta forma se dificultan los intentos de ingeniería inversa y desensamblado que tienen la intención de obtener una forma de código fuente cercana a la forma original [Wikipedia, 2021].

2.2.3. Métodos de ofuscación de código fuente

[Novak *et al.*, 2019] Explica que existen muchos métodos de ofuscación utilizados por estudiantes para ocultar la similitud, a su vez menciona que en un estudio de 72 artículos se identificaron 25 métodos de ofuscación, y de estos se especificaron 16 métodos distintos. A continuación se mencionan algunos métodos de ofuscación de código fuente:

- [Marzieh *et al.*, 2011] Mencionan cambios en el formato del código, como la modificación de espacios en blanco como sangrías, espacios, nuevas líneas, etc.
- [Marzieh *et al.*, 2011] Mencionan cambios en los comentarios del código.
- [Đurić y Gašević, 2012] y [Donaldson *et al.*, 1981] Mencionan el cambio de los nombres de los identificadores. como nombres de variables, nombres de constantes, nombres de funciones, nombres de clases, etc.
- [Donaldson *et al.*, 1981] Menciona cambios en el orden de las declaraciones de la variables.
- [Grier, 1981] Menciona agregar líneas de código redundantes, líneas de código que no hacen nada.
- [Donaldson *et al.*, 1981] Menciona dividir una línea de código en varias líneas.
- [Whale, 1990] Menciona el reemplazo de la llamada de un procedimiento por el procedimiento.

- [Whale, 1990] Menciona el cambio de la especificacion de una declaracion como el cambio de los operaciones y el operando, alternando modificadores cambio de tipos de datos.
- [Marzieh *et al.*, 2011] Mencionan el cambio de estructuras de control equivalentes, como el reemplazo de estructuras repetitivas y condicionales.
- [Đurić y Gašević, 2012] Mencionan la simplicacion de codigo, eliminar lineas de codigo no necesarias.

2.3. Conceptos de compiladores

2.3.1. Procesador de lenguaje

Un procesador de lenguaje se muestra en la figura 2.1. También llamado compilador es un programa que puede leer un programa en un lenguaje y traducirlo en un programa equivalente en otro lenguaje. Una función importante del compilador es reportar cualquier error en el programa fuente que detecte durante el proceso de traducción [Aho, 2008].

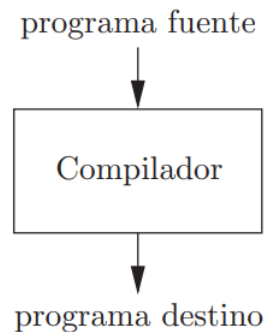


Figura 2.1: Un compilador
Nota: Figura tomada de [Aho, 2008].

El proceso de compilación opera como una secuencia de fases con las cuales transforma un programa fuente en otro, se muestra en la figura 2.2.

En el presente trabajo de investigación, solo es de interés de cubrir los teoría y conceptos hasta el análisis semántico.

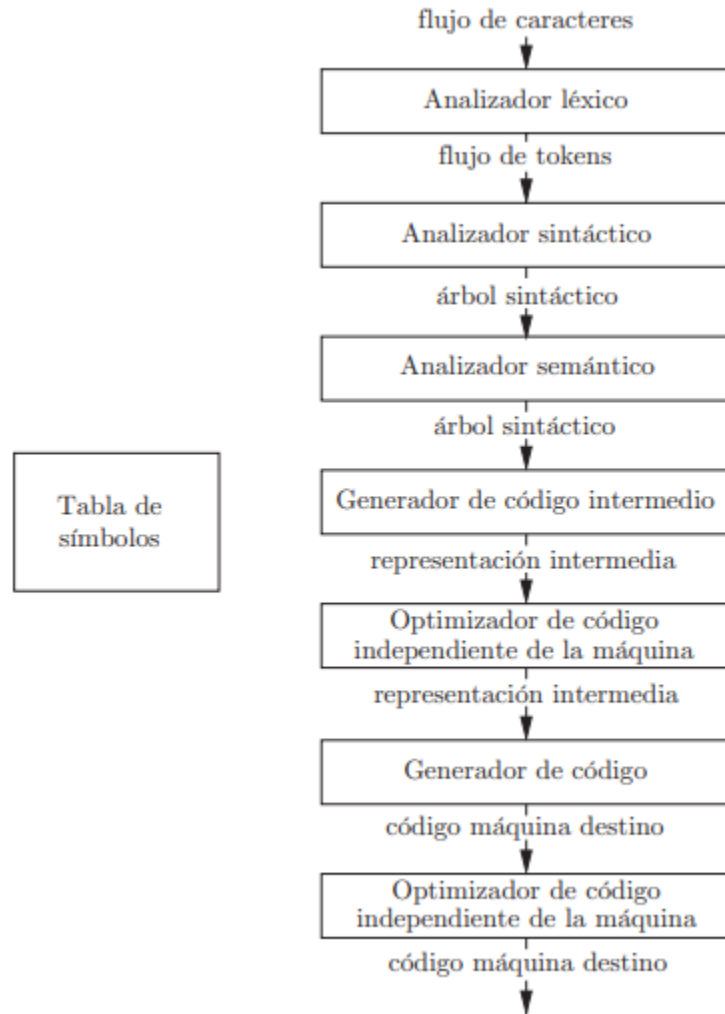


Figura 2.2: Fases de un compilador
 Nota: Figura tomada de [Aho, 2008].

2.3.2. Análisis Léxico

[Aho, 2008] Explica que la primera fase de un compilador se le llama análisis léxico o escaneo. El analizador léxico lee el flujo de caracteres que componen el programa fuente y los agrupa en secuencias significativas, conocidas como lexemas. Para cada lexema, el analizador léxico produce como salida un token. Estos lexemas son los que pasaran a la siguiente fase, el análisis de la sintaxis. El analizador léxico ignora los espacios en blanco que separan los lexemas.

[Catalán, 2010] Explica que esta fase consiste en leer el texto del código fuente

carácter a carácter e ir generando los tokens. Estos tokens constituyen la entrada para el siguiente proceso de análisis. El agrupamiento de caracteres en tokens depende del lenguaje que vayamos a compilar. Es decir un lenguaje generalmente agrupara caracteres en tokens diferentes de otro lenguaje. Los tokens pueden ser de dos tipos, cadenas específicas como palabras reservadas, puntos y comas, etc., y no específicas, como identificadores, constantes y etiquetas. La diferencia entre ambos tipos de tokens radica en si ya son conocidos previamente o no. El analizador léxico irá ignorando las partes no esenciales para la siguiente fase, como pueden ser los espacios en blanco, los comentarios, etc., es decir, realiza la función de preprocesador en cierta medida. Por lo tanto, y resumiendo, el analizador léxico lee los caracteres que componen el texto del programa fuente y suministra tokens al analizador sintáctico.

2.3.3. Análisis Sintáctico

[Aho, 2008] Explica que la segunda fase del compilador es el análisis sintáctico o parsing. El parser utiliza los primeros componentes de los tokens producidos por el analizador de léxico para crear una representación intermedia en forma de árbol que describa la estructura gramatical del flujo de tokens. Una representación típica es el árbol sintáctico, en el cual cada nodo interior representa una operación y los hijos del nodo representan los argumentos de la operación. En la figura 2.3 se muestra un árbol sintáctico para el flujo de tokens como salida del analizador sintáctico.

[Catalán, 2010] Explica que el analizador sintáctico tiene como entrada los lexemas que le suministra el analizador léxico y su función es comprobar que están ordenados de forma correcta dependiendo del lenguaje que se quiere procesar. Los dos analizadores suelen trabajar unidos e incluso el léxico suele ser una subrutina del sintáctico. El analizador sintáctico se le suele llamar parser, este genera de manera teórica un árbol sintáctico. Este árbol se puede ver como una estructura jerárquica que para su construcción se usa reglas recursivas. La estructuración de este árbol hace posible diferenciar entre aplicar unos operadores antes de otros en la evaluación de expresiones. En resumen la tarea del analizador sintáctico es procesar los lexemas que le suministra el analizador léxico, comprobar si están bien ordenados, y si no lo están, generar los informes correspondientes. Si la ordenación, se

generará un árbol sintáctico teórico.

2.3.4. Análisis Semántico

[Aho, 2008] Explica que el analizador semántico utiliza el árbol sintáctico y la información en la tabla de símbolos para comprobar la consistencia semántica del programa fuente con la definición del lenguaje. También recopila información sobre el tipo y la guarda, ya sea en el árbol sintáctico o en la tabla de símbolos, para usarla más tarde durante la generación de código intermedio. Una parte importante del análisis semántico es la comprobación de tipos, en donde el compilador verifica que cada operador tenga operando que coincidan.

[Catalán, 2010] Explica que esta fase toma el árbol sintáctico teórico de la anterior fase y hace una serie de comprobaciones antes de obtener un árbol semántico teórico. Esta fase es quizás la más compleja. Hay que revisar que los operadores trabajan sobre tipos compatibles, si los operadores obtienen como resultado elementos con tipos adecuados, si las llamadas a subprogramas tienen los parámetros adecuados tanto en número como en tipo, etc. Esta fase debe preparar el terreno para atajar las fases de generación de código y debe lanzar los mensajes de error que encuentre. En resumen, su tarea es revisar el significado de lo que se va leyendo para ver si tiene sentido.

2.3.5. Árboles de sintaxis abstracta

[Aho, 1999] Explica que un árbol es una estructura jerárquica sobre una colección de objetos, como ejemplos están los arboles genealógicos y organigramas. Los arboles son útiles para analizar circuitos eléctricos y para representar estructuras de formulas matemáticas. Dentro de la área de la computación se usan para organizar la información en sistemas de bases de datos y para representar la estructura sintáctica de un programa fuente en compiladores.

En [Cormen *et al.*, 2009] podemos obtener una definición y propiedades matemáticas de un árbol, para ello se desarrollaran algunas definiciones de teoría de grafos. Sea G un grafo no direccionado es representado por un par $G = (V, E)$, donde V representa el conjunto de vértices de G , se le denota con un círculo. E representa el conjunto de aristas

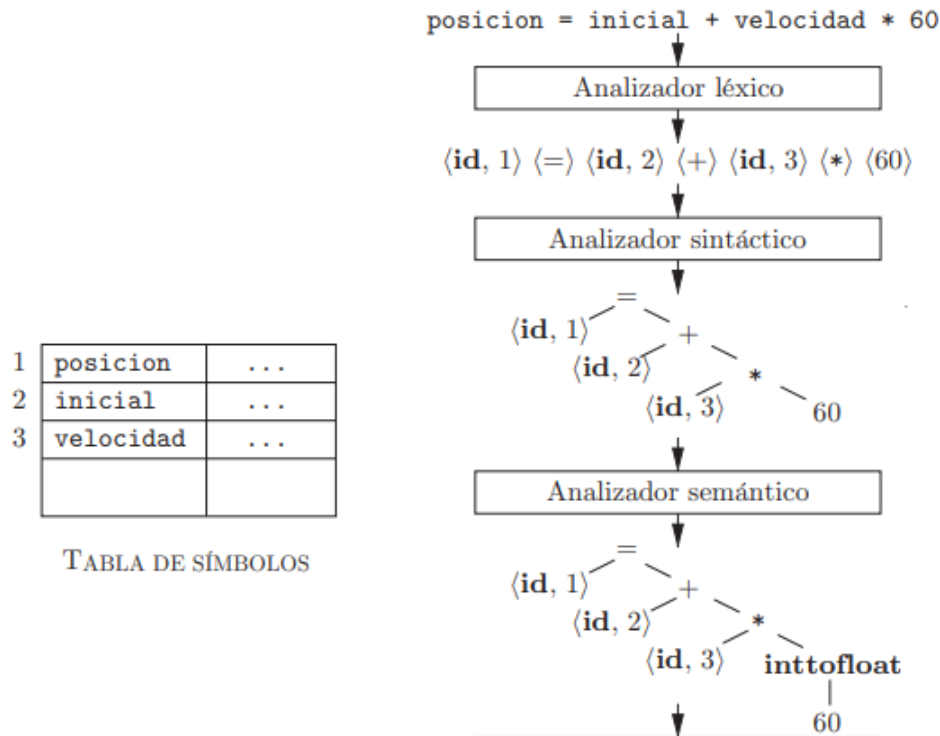


Figura 2.3: Traducción de una asignación
 Nota: Figura tomada de [Aho, 2008].

de G , se le denota con una línea.

Además para que un grafo sea considerado un árbol deben cumplirse las siguientes afirmaciones:

- Dos vértices de G están conectados por un camino único.
- G es conexo, pero si se elimina cualquier vértice que no sea una hoja de E , el grafo resultante es no conexo.
- G es conexo, y $|E| = |V| - 1$.
- G es acíclico, y $|E| = |V| - 1$.
- G es acíclico, pero si cualquier vértice es agregado a E , el grafo resultante contendrá un ciclo.

En resumen un árbol es un grafo no direccionado, conexo, acíclico, y que dos vértices cualesquiera están conectados por un camino único, como se muestra en la figura 2.4.



Figura 2.4: Árbol

Nota: Figura tomada de [Cormen *et al.*, 2009].

Las estructuras de datos basadas en árboles se hicieron muy populares en el campo del desarrollo de compiladores: cada vez que se envía un archivo que contiene el código fuente al compilador, se realizan varios pasos antes de que se puedan generar las instrucciones de la máquina. El código tiene que ser tokenizado por un Lexer primero: separa el flujo de entrada en tokens individuales y los pasa al analizador, que utiliza una gramática libre de contexto del lenguaje de programación para construir una representación de código intermedio, un llamado árbol de análisis. Cada token encontrado por el lexer está representado por un nodo en el árbol de análisis. No todos los tokens (nodos) tienen un valor semántico: algunos tokens, por ejemplo, paréntesis y punto y coma, son puramente sintácticos. Por lo tanto, puede omitirse. La estructura de datos resultante se denomina árbol de sintaxis abstracta (AST). Ahora que el código fuente se representa como un árbol, se puede analizar de una forma más sofisticada manera que al analizar el flujo de token plano: el árbol se puede recorrer o buscar de varias maneras por ejemplo, pos-orden, pre-orden, búsqueda profundidad, etc. [Würsch *et al.*, 2022].

2.4. Conceptos de algorítmica

2.4.1. Programación dinámica

La programación dinámica es una técnica matemática que se utiliza para la solución de problemas matemáticos seleccionados, en los cuales se toma una serie de decisiones en forma secuencial.

2.4.2. Distancia de edición de árbol

La distancia edición o distancia entre palabras es el número mínimo de operaciones requeridas para transformar una cadena de caracteres en otra, se usa ampliamente en teoría de la información y ciencias de la computación. Se entiende por operación, bien una inserción, eliminación o la sustitución de un carácter. Esta distancia recibe ese nombre en honor al científico ruso Vladimir Levenshtein, quien se ocupó de esta distancia en 1965. Es útil en programas que determinan cuán similares son dos cadenas de caracteres, como es el caso de los correctores ortográficos.

La distancia de edición del árbol se define como la secuencia de costos mínimos de operaciones de edición de nodos que transforman un árbol en otro. Las operaciones consisten en eliminar, agregar y modificar nodos del árbol.

2.4.3. Conjuntos Disjuntos

En [Cormen *et al.*, 2009] obtenemos una la definición de conjuntos disjuntos y la estructura de datos para conjuntos disjuntos.

Los conjuntos disjuntos es una colección representada por $S = \{S_1, S_2, \dots, S_k\}$ donde cada elemento S_i representa un conjunto dinámico, y para cada elemento de la colección se cumple que $S_i \cap S_j = \emptyset$.

La estructura de datos para conjuntos disjuntos soporta eficientemente las operaciones de crear un conjunto, búsqueda del representante de un conjunto y la unión de conjuntos.

Sean u y v elementos de un conjunto:

- **MAKE – SET(u)** Crea un nuevo conjunto cuyo único miembro es u . Dado que los conjuntos son disjuntos, es necesario que u no se encuentre en otro conjunto.
- **UNION(u, v)** Une los conjuntos dinámicos que contienen a u y v , es decir $S_u \cap S_v$, el representante del conjunto resultante sera cualquiera de los dos.
- **FIND – SET(u)** devuelve un puntero al representante del conjunto que contiene a u .

2.5. Herramientas para la detección de similitud de código

2.5.1. Algoritmos utilizados para la detección de similitud

[Novak *et al.*, 2019] Identificó diferentes algoritmos a continuación se mencionan algunos de ellos:

- Recuento de atributos
- Huella digital
- Coincidencia de cadena
- Texto base
- Estructura base
- Estilo
- Semántico
- N-gramas
- Árboles
- Grafos

Algunos de estos fueron inventados en la década de 1980. También hace mención a que los enfoques basados en estructuras son mucho mejores y que la mayoría de las herramientas de detección de similitud combinan más de un tipo de algoritmo.

2.5.2. Detección léxica

Las técnicas y herramientas para computar las diferencias textuales entre documentos son bien conocidas y aprobadas. Sin embargo, las herramientas existentes como GNU diff tratan con información plana, en lugar de jerárquica. Por lo general, calculan una lista de líneas que deben cambiarse, insertarse o eliminarse para que un primer documento coincida con el segundo. [Würsch *et al.*, 2022].

Bibliografía

- [Aho, 2008] Aho, A. (2008). *Compiladores: principios, tecnicas y herramientas*. Pearson Educacion, Mexico.
- [Aho, 1999] Aho, A. V. (1999). *Estructuras de datos y algoritmos*. Addison Wesley Longman, Singapore, Singapore.
- [Catalán, 2010] Catalán, J. (2010). *Compiladores : teoría e implementación*. RC Libros, San Fernando de Henares, Madrid.
- [Cheers *et al.*, 2021] Cheers, H., Lin, Y., y Smith, S. (2021). Academic source code plagiarism detection by measuring program behavioural similarity.
- [Cormen *et al.*, 2009] Cormen, T. H., Leiserson, C. E., Rivest, R. L., y Stein, C. (2009). *Introduction to Algorithms*. The MIT Press. MIT Press, London, England.
- [Donaldson *et al.*, 1981] Donaldson, J. L., Lancaster, A.-M., y Sposato, P. H. (1981). A plagiarism detection system. pp. 21–25.
- [Díaz *et al.*, 2007] Díaz, J., Banchoff, L., y Rodríguez, L. (2007). "herramientas para la detección de plagio de software. un caso de estudio en trabajos de cátedra".
- [Grier, 1981] Grier, S. (1981). A tool that detects plagiarism in pascal programs. En *Proceedings of the Twelfth SIGCSE Technical Symposium on Computer Science Education*, SIGCSE '81, p. 15–20, New York, NY, USA. Association for Computing Machinery.
- [Hage y Rademaker, 2010] Hage, J. y Rademaker, P. (2010). A comparison of plagiarism detection tools.

- [Marzieh *et al.*, 2011] Marzieh, A., Mahmoudabadi, E., y Khodadadi, F. (2011). Pattern of plagiarism in novice students generated program: An experimental approach. *The Journal of Information Technology Education*, 10.
- [Novak *et al.*, 2019] Novak, M., Joy, M., y Kermek, D. (2019). Source-code similarity detection and detection tools used in academi: A systematic review. *ACM Trans. Comput. Educ.*, 19(3).
- [Pachón, 2019] Pachón, H. (2019). *Generación de un algoritmo para el análisis de similitudes de código fuente en lenguajes Java y Python*. Uniandes.
- [Whale, 1990] Whale, G. (1990). Software metrics and plagiarism detection. *Journal of Systems and Software*, 13(2):131–138. Special Issue on Using Software Metrics.
- [Wikipedia, 2021] Wikipedia (2021). Obfuscation (software) — Wikipedia, the free encyclopedia. [http://en.wikiipedia.org/w/index.php?title=Obfuscation%20\(software\)&oldid=1056756228](http://en.wikiipedia.org/w/index.php?title=Obfuscation%20(software)&oldid=1056756228). [Online; accessed 02-December-2021].
- [Würsch *et al.*, 2022] Würsch, M., Gall, H., Fluri, B., y Kiefer, C. (2022). Improving changedistiller improving abstract syntax tree based source code change detection.
- [Đurić y Gašević, 2012] Đurić, Z. y Gašević, D. (2012). A Source Code Similarity System for Plagiarism Detection. *The Computer Journal*, 56(1):70–86.