

Detección automática de plagio basada en la distinción y fragmentación del texto reutilizado

Por:

José Fernando Sánchez Vega

Tesis sometida como requisito parcial para obtener el grado de:
Maestro en Ciencias en el Área de Ciencias Computacionales
en el Instituto Nacional de Astrofísica,
Óptica y Electrónica.

Supervisada por:

Dr. Luis Villaseñor Pineda

Comité revisor:

Dr. Aurelio López López
Dr. Francisco Trinidad Martínez
Dr. Saúl E. Pomares Hernández

© INAOE Enero 2011

El autor otorga al INAOE el permiso de reproducir y distribuir copias
en su totalidad o en partes de esta tesis



Detección automática de plagio basada en la distinción y fragmentación del texto reutilizado

Tesis de Maestría

Por:

José Fernando Sánchez Vega

Asesor:

Dr. Luis Villaseñor Pineda

Comité revisor:

Dr. Aurelio López López (INAOE)

Dr. Francisco Trinidad Martínez (INAOE)

Dr. Saúl E. Pomares Hernández (INAOE)

Instituto Nacional de Astrofísica, Óptica y Electrónica,
Coordinación de Ciencias Computacionales
Luis Enrique Erro # 1, Tonantzintla,
Puebla, 72840, México
25 de enero de 2011

RESUMEN

En el panorama actual existen gran cantidad de documentos digitales que pueden ser fácilmente consultados; estas grandes bibliotecas (llámense bibliotecas virtuales o la Internet pública) contienen obras que abarcan una gran variedad de temas con una enorme diversidad de enfoques. Al mismo tiempo de este apogeo de información “facil”, se está dando un nuevo auge de la reutilización, y el problema es que esta reutilización es inescrupulosa; se realiza sin dar cuenta de que los contenidos provienen de obras originales, lo que aleja el material de discusión de sus verdaderos autores, sin darles el crédito correspondiente. Estos “abusos” de la información constituyen un robo de material intelectual conocido como plagio.

La detección de plagio es la respuesta natural al desequilibrio que han generado las tecnologías de la información frente a los autores que se mantienen produciendo materiales originales. Es importante atender a estos sectores, pues es aquí donde se lleva a cabo la producción y comunicación del conocimiento.

En la detección automática de plagio (DAP), un documento (el cual se sospecha presenta alguna clase de plagio o del cual se quiere comprobar que no exista posibilidad alguna de contener plagio) es comparado de manera automática por una computadora con alguna fuente particular para evaluar si se trata de un plagio.

Las técnicas de la DAP típicamente realizan la detección midiendo la cantidad de texto compartido entre los dos documentos; la principal dificultad en esta tarea es que no todo el texto compartido se debe a un plagio, pues la existencia de coincidencias temáticas o de estilo produce porciones de texto común sin ser necesariamente plagio. Para atacar esta dificultad proponemos una representación con atributos que capturan la fragmentación y la distinción del texto compartido. La fragmentación del texto compartido es capturada utilizando una serie de atributos que contabilizan las secuencias de texto compartido especializándose cada atributo en una longitud particular. La distinción del texto compartido es utilizada para ponderar cada una de las secuencias compartidas; esta

ponderación mide tanto la relevancia temática como la usabilidad (que tanto fue usado por el posible plagiario) del texto compartido.

También, para poder afrontar el problema del texto plagiado que es modificado para evitar su detección, se propone un novedoso modelo que permite aumentar el porcentaje de detección del texto que fue tomado del documento original. Se utilizan criterios flexibles en la búsqueda del texto utilizado, contemplando no sólo la copia exacta sino que también algunas posibles modificaciones. Esta innovación, dota al método de detección de evidencia más completa para poder decidir sobre el plagio.

Estos atributos, que caracterizan al texto común, al utilizarse con algoritmos de aprendizaje automático, brindan más elementos para detectar el plagio de una manera más eficaz. En experimentos con los pocos *corpus* que existen de plagio, hemos obtenido una mejora relativa de 7% en la precisión general. También observamos que el método propuesto presenta menor sensibilidad al cambio del dominio con respecto a otros métodos.

ABSTRACT

In the current scenario there are many digital documents which are “easily” accessed; these big libraries (whatever you call them, virtual libraries or the public Internet) contain works covering a wide variety of topics with a huge diversity of approaches. At the same time in the apogee of "easy" information, it is being a resurgence of reuse, and the problem is that this reuse is unconscionable; it is done without notice that the content come from another original works and without the corresponding credit. These "misuses" of information constitute a theft of intellectual material known as plagiarism.

The detection of plagiarism is the natural response to the imbalance that generated the information technology, which does not protect the authors, who remain producing original material. It is important to address these areas, because this is where it carries out the production and communication of knowledge.

In the automatic plagiarism detection (APD) a document (which presents a kind of suspected plagiarism or which you want to check that there is no possibility of containing plagiarism) is automatically compared by a computer with a particular source to assess whether it is plagiarism.

The APD techniques typically perform the detection by measuring the amount of shared text documents between two documents; the main difficult in this task is that not all the shared text is due by a plagiarism; because the existence of thematic or stylistic similarities produces lots of common text but not necessarily plagiarism. To attack this difficulty we propose a representation with attributes that capture the fragmentation and allow to make a distinction of the common text. The fragmentation of the common text is captured using a series of attributes that account for common text strings; each of these attributes is specialized in a particular string length. The distinction of the common text is used to weigh each of the common sequences, this weight schema measures both the thematic relevance and usability (how much of the text was used by the potential plagiarist) of common text.

Also, to deal with the problem of plagiarized text, that is modified to avoid detection, a new model is proposed which increases the detection rate of the text that was taken from the original document. We improve certain flexible criteria which are used in the search of the reused text, considering not only the exact copy but also some possible modifications. This innovation provides to the detection method more complete evidence in order to decide about when plagiarism happens.

When these attributes, which characterize the common text, are used with machine learning algorithms, provide more elements to detect more effectively the plagiarism. In experiments with the few existing corpus of plagiarism, we obtained a relative improvement of 7% in overall accuracy. We also note that the proposed method is less sensitive to the change of the domain with respect to other methods.

Agradecimientos

Empezaré agradeciendo a mis asesores Luis Villaseñor Pineda y Manuel Montes y Gomes. A Manuel, que sin tener el reconocimiento institucional de ser mi asesor, siempre estuvo, al igual que Luis, dispuesto a discutir los resultados de los experimentos y a recomendar los mejores caminos a seguir. A Luis, que además me ayudó a seleccionar las mejores palabras (en lo posible) para expresar el proceso de un año de investigación.

Quiero agradecer a los sinodales por participar en el proceso de corrección, ayudando a que cualquier posible futuro lector pueda entender mejor el trabajo que realicé.

Agradezco a mis padres, quienes siempre me apoyaron y quienes, aunque poco sabían del tema, cedieron parte de su tiempo para darme algunas primeras correcciones de la tesis y ayudar a que la tesis fuera un poco más amena y fácil de leer.

De igual modo quiero agradecer a Ruth, mi compañera en la vida, quien siempre me animó en aquellos malos momentos en los que los resultados de los experimentos no eran buenos y con quien disfrute, en gran medida, compartir los buenos momentos.

Por último quiero agradecer a mi pequeña y fiel acompañante de desveladas de trabajo, mi perrita Orientina, quien aunque se caía de sueño y se moría de ganas de ir a un lugar más cómodo, permanecía a mi lado hasta que terminara mi trabajo (y quien aún está aquí a mi lado).

Finalmente quiero decirles a todos los involucrados y acompañantes en el proceso de esta tesis y de mi maestría, que fue una experiencia enriquecedora y muy feliz. Gracias por estar a mi lado en este camino.

Índice

Resumen	V
Abstract	VII
Agradecimientos	IX
1. Introducción	1
1.2 La detección del plagio como un problema computacional	3
1.3 Objetivos	5
1.4 Contenido de la tesis	5
2. Estado del arte	7
2.1 Nacimiento de los sistemas de detección automática de plagio	7
2.2 Los rasgos observados en la detección automática de plagio	8
2.3 La problemática particular de la detección automática de plagio	11
2.4 Trabajos previos en la detección automática de plagio	14
2.4.1 Enfoques Simples	15
2.4.2 Enfoques estructurados dependientes de RL	17
2.4.3 Análisis estructural independiente de RL	18
2.5 Los rasgos observados en los diferentes métodos de DAP	23

2.6	Decisión de plagio en la DAP	25
3.	El método propuesto	29
3.1	Introducción al método	29
3.1.1	Secuencias comunes	31
3.2	Atributos de fragmentación	33
3.3	Atributos de distinción	36
3.4	Generalización del concepto del texto reutilizado	40
3.4.1	Índice de Reescritura	41
3.4.2	Algoritmo de búsqueda para la obtención del IRe	45
3.4.3	Las cadenas de texto reutilizado	49
3.4.4	Los atributos de las cadenas de texto reutilizado	50
4.	Evaluación	53
4.1.1	Disponibilidad de los <i>corpus</i> de Plagio.....	53
4.1.2	<i>Corpus</i> METER	54
4.1.3	<i>Corpus Plagiarised Short Answers</i>	55
4.2	Métricas de evaluación	56
4.3	Condiciones de los experimentos	59
4.3.1	El clasificador	59
4.3.2	El método de evaluación	60
4.3.3	Preprocesamiento del <i>corpus</i>	60
4.4	Métodos de referencia	61
4.5	Resultados de los métodos de referencia	63
4.6	Discusión de los métodos de comparación	67
4.7	Experimentos del método propuesto	70
4.8	Resultados de los métodos propuestos	73

4.9	Discusión de los resultados	74
5.	Conclusión	77
5.1	Recapitulación	77
5.2	Conclusiones	78
5.3	Trabajo futuro	79
6.	Referencias	80
A.	Apéndice I	89
B.	Apéndice II	95
C.	Apéndice III	99

1. Introducción

Existen múltiples definiciones formales del plagio¹, aunque es difícil nombrar una como la más acertada [1]; en lo que coinciden todas estas definiciones es en que:

“El plagio es tomar ideas de otros y presentarlas sin dar el crédito correspondiente al autor original”.

En la actualidad, los casos de plagio han aumentado debido a la facilidad de encontrar y reutilizar contenidos de obras previamente escritas². Basada en estas prácticas, una descripción pragmática es: se está cometiendo la falta ética conocida como plagio cuando la información es ocupada simplemente copiándose sin ningún estricto procesamiento “racional” y con pocas o nulas adaptaciones a un contexto [2].

La gran cantidad de herramientas del procesamiento, gestión, descubrimiento y filtrado de información han permitido aprovechar el enorme cúmulo de obras, libros,

¹ En todo este documento nos referimos únicamente al plagio que sucede en documentos de texto, dejando fuera el plagio de otras obras intelectuales como logos, fotografías, pinturas, filmes, etc.

² El conocido “*copy-paste*” permite extraer porciones de texto de fuentes originales e incorporarlas a nuevos documentos, si no se da el crédito y referencia de estos extractos, así como el entrecorillado de las porciones de cita textuales, entonces, se está cometiendo plagio.

coleccionos y datos, existentes en las millones de fuentes disponibles a través de las bases de datos de artículos científicos y de la WWW [3].

Las nuevas facilidades de acceso a la información (en muchos casos de forma gratuita, incluso para publicaciones científicas de gran prestigio) crean una plataforma formidable para los estudiantes y para muchos investigadores en la que pueden fortalecer su formación y enriquecer sus conocimientos. Lamentablemente, estas facilidades han provocado, en escuelas y universidades, la presentación de algunos trabajos poco escrupulosos, que son conformados al plagiar materiales previamente publicados, convirtiéndose en un problema serio para la educación [4].

Adicionalmente a los materiales en Internet que son indebidamente robados por los estudiantes para conformar sus trabajos escolares, existen varias paginas como: www.schoolsucks.com, www.cheathouse.com, www.seminarky.cz, www.monografias.com, www.rincondelvago.com, etc., que cuentan con miles de trabajos adecuados a diferentes necesidades académicas [5] y motivan el uso de estos contenidos para reconstruir versiones con mínimas modificaciones para ser presentados como si fuera un trabajo nuevo u original.

En el ámbito de la investigación científica, el problema también va en aumento; como se afirma en [6] la recepción de trabajos plagiados ha estado extendiéndose en los comités de publicación de revistas prestigiosas.

No todos los casos de plagio han podido ser identificados antes de que se permitiera su publicación. En el *corpus* de artículos ya publicados, MedLine, la NLM (*National Library of Medicine*) declaró haber encontrado 607 “publicaciones duplicadas” y cabe aclarar que la NLM considera a las “publicaciones duplicadas” como artículos que presentan prácticamente la misma información, es decir, serían aquellas en las que todas las secciones son plagiadas de un mismo artículo (una única fuente del plagio) [7].

Existen códigos de honor [8] que velan por la correcta ética de las publicaciones. Inclusive la Política de Ciencia y Tecnología de Estados Unidos (*Office of Science and Technology Policy*) condena el plagio y lo considera tan grave como la fabricación o falsificación de resultados o información [7].

En toda esfera, el plagio es una actividad indeseable; la intención de los trabajos escolares es que los alumnos puedan construir sus ideas u opiniones y las puedan presentar y defender con argumentos a partir de sus lecturas, habilidades que no se desarrollarán si se presentan trabajos que no han sido realizados por ellos. De igual manera, en la investigación científica, el plagio impacta en la calidad de los *corpus* de artículos científicos, pues merma la producción. Para un investigador, leer las mismas ideas, incluso con redacciones similares, es agotante y hace crear un falso panorama de la falta de diversidad de propuestas en cierta área de interés [7].

Las herramientas de manejo de información (filtrado, búsqueda, etc.) son sumamente útiles para satisfacer necesidades específicas e incluso para obtener panoramas generales de las obras existentes en campos o disciplinas particulares. Sin embargo, estas herramienta han resultado ser (como se indica en [9]) “armas de doble filo”, al permitir el sencillo acceso a la información, pero, a su vez, facilitar la inapropiada utilización de estos contenidos. Como Birn Sergey indica “La tecnología actual no mantiene un buen balance entre protección [...] y acceso a la información” [10]. La tarea de la detección automática de plagio nace con la motivación de crear herramientas de igual complejidad y potencia que las disponibles para el acceso de la información, pero que permitan custodiar la integridad de las nuevas creaciones (para mantener la accesibilidad existente).

1.2 La detección del plagio como un problema computacional.

En [1] se discute acerca del nivel en el que ocurre el plagio. Existen dos niveles: el de las ideas y el de las palabras. La detección automática se limita a tomar las decisiones con base en los textos y, por tanto, está enfocada principalmente al plagio de las palabras debido a las limitaciones computacionales en el entendimiento de las ideas. Pero no perdamos la perspectiva de que las ideas deben de ser planteadas de igual manera en palabras y por lo tanto generarán textos que son susceptibles de ser analizados por los sistemas de detección. En particular, el plagio de ideas se podría poner en la cima de la jerarquías de

reformulación, en donde, se podría tratar como una paráfrasis si la forma de reexpresar las ideas no es en extremo diferente a como se ha planteado en el original.

El principal fenómeno de interés en la detección automática de texto es la reutilización de contenidos [11] o la también llamada reutilización de texto. El plagio en este contexto es cuando se reutiliza texto ilegítimamente, sin permiso y sin dar el crédito correspondiente [12]. El texto reutilizado, como se ha dicho antes, puede tener cambios o incluso puede ser una reescritura libre; la distancia que separa al texto original de su plagio depende de la cantidad y complejidad de las modificaciones, que pueden llevar al texto desde una copia, un resumen o paráfrasis, hasta una simple relación temática [13]. Los cambios que sufre el texto al ser reutilizado han sido modelados por algunos sistemas, como operaciones sobre las palabras u oraciones: la eliminación o inserción de palabras, el cambio de palabras (principalmente por sinónimos) y el reordenamiento de palabras (debido a la reescritura de oraciones, como es el cambio de voz activa a pasiva) [14,15].

Los nuevos sistemas de detección han enfocado su atención en cuantificar el texto reutilizado, esperando que las operaciones de reescritura no deterioren sus medidas de la reutilización. Con estas medidas intentan cuantificar el monto de texto reutilizado (según sus sistemas) necesario para que un documento sea considerado plagio; desafortunadamente, el monto de material común entre dos trabajos, para que se consideren como plagio, depende del área del conocimiento en que se realiza la aportación [16,17]. Así, por ejemplo, los trabajos de matemáticas pueden contener muchos párrafos copiados de forma literal, debido a la necesaria enunciación de los teoremas que fundamentan la propuesta; por otro lado, los países aún no han acordado una legislación estándar que defina la cantidad de texto común que debe de tener un par de documentos para que se consideren plagiados [18].

Debido a que las características de los casos de plagio varían según el área del conocimiento y según lo dispuesto por las legislaciones de cada país, es deseable que los sistemas de detección automática de plagio puedan aprender de forma automática estas características. Por esto, nuestra propuesta permite, a partir de la caracterización de ejemplos de plagio en un dominio específico, aprender modelos que les permiten detectar el plagio en otros documentos.

1.3 Objetivos

El objetivo general de esta investigación es:

- Proponer y obtener un método de detección de plagio que aprenda a distinguirlo a partir de ejemplos previamente proporcionados, utilizando técnicas de aprendizaje automático.

Los objetivos particulares que se desprenden son:

- Obtener atributos que describan características del texto común entre los documentos.
- Obtener el estado de las herramientas existentes más efectivas para capturar el texto común.
- Obtener un marco de evaluación que permita comparar la efectividad del método propuesto con los principales métodos modernos de la detección de plagio.

1.4 Contenido de la tesis

Además de este capítulo introductorio el resto del documento se encuentra organizado de la siguiente manera:

Capítulo 2: En éste se expone el estado del arte de la detección automática de plagio. Se explica cómo es que se ha realizado esta tarea, tradicionalmente llevada a cabo de forma manual, utilizando la computadora. Se abordan algunas de las estrategias utilizadas por los expertos que han inspirado algunos de los mecanismos usados en la detección automática de plagio. En esta sección, se determinan las características propias de la tarea de la detección automática de plagio y se explica la diferencia de algunas otras tareas hermanas en el análisis automático del plagio. Finalmente se presenta de forma organizada y esquemática las diversas propuestas encontradas en la literatura, englobadas en dos enfoques cuya principal diferencia

es la utilización de recursos adicionales y se introduce brevemente las características del estado del arte que generaron las innovaciones presentadas en esta tesis.

Capítulo 3: Es la parte medular del trabajo, es en donde se presenta el método propuesto.

En la primera sección de este capítulo se introducen brevemente las principales novedades presentes en el método así como algunas de las motivaciones que llevaron a proponerlas. En el resto del capítulo se abordan de manera profusa las características y la forma en la que el método actúa sobre los casos en los que se sospecha de plagio. A lo largo de estas explicaciones, en cada cierre de sección se presenta cómo es que el método está conformado hasta ese punto, esto se realiza con el fin de dar más claridad a cada elemento del método así como predisponer al lector a esperar las evaluaciones parciales que se realizan en la siguiente sección.

Capítulo 4. En este capítulo se expone la forma en que se evaluó el método propuesto.

Primero se exponen las principales características de los *corpus* utilizados así como las razones por las que se eligió utilizarlos. Después se presentan los métodos de referencia con los que se va a comparar el método propuesto, así como las motivaciones por las que se eligieron (y crearon) estos métodos de referencia, de igual manera se exponen y discuten los resultados obtenidos en su evaluación. Se presentan los experimentos con los que se evaluará el método propuesto así como cada una de las innovaciones que éste contiene. Luego se muestran los resultados de los experimentos que lo evalúan, esto se discute y se compara con los de referencia.

Capítulo 5. En este capítulo se expone una breve recapitulación de todo el trabajo realizado

y se enlistan las principales conclusiones a las que se ha llegado en la investigación realizada. La sección titulada “Trabajo futuro”, expone las principales posibilidades que se abren para poder continuar el trabajo aquí propuesto, ahí se manifiestan las múltiples opciones en varias ramas o direcciones distintas.

2. Estado del arte

2.1 Nacimiento de los Sistemas de detección automática de plagio.

Ante la imposibilidad de los organismos (comisiones éticas y universidades) para poder evitar el plagio y a la abrumadora cantidad de sitios que deben ser vigilados (editoriales de revistas, sitios web y las aulas escolares), fue que nacieron los primeros sistemas de detección automática de plagio. La universidad de Virginia desarrollo el WcopyFind, la universidad de Hertfordshire propuso el Ferret Plagiarism Detector y Stanford su SCAM (Stanford Copy Analysis Mechanism).

Como se muestra en [19], muchas de estas primeras soluciones no mostraron muy buena eficacia; esto se debía principalmente a cualquiera de las siguientes dos razones: (i) Algunas de las propuestas eran técnicas no muy apropiadas para esta tarea pues estaban muy influenciadas por las preexistentes de la recuperación de información; o (ii) El abuso de la tendencia por usar técnicas de *fingerprint*. Estas técnicas eran utilizadas principalmente por las librerías *on-line* que tenían interés en saber si alguno de los contenidos que comercializaban se encontraban disponibles en la red (colocados

posiblemente por algún inescrupuloso cliente), por tanto, las técnicas con este enfoque sólo han sido efectivas para encontrar plagios en los cuales el texto robado presente pocas o nulas modificaciones.

Algunos de estos primeros sistemas universitarios proporcionaban análisis que, como se expone en [20, 21], sólo servían para indicarle a una persona (habitualmente el profesor) en qué documentos debía presentar mayor atención, requiriendo después de obtener los resultados del sistema, una confirmación manual del análisis computacional y, posteriormente, una investigación para cada caso que el operador considerara altamente sospechoso. Los sistemas actuales aún no pueden ser absolutamente autónomos (pues su exactitud no es del 100%), pero las nuevas técnicas permiten emitir de forma automática un juicio preliminar de la existencia de plagio e inclusive algunos pocos pueden indicar el tipo de plagio.

Entre más elaborado sea el plagio más difícil es que los sistemas automáticos lo puedan detectar; inclusive algunos estudiantes han aprendido el tipo y cantidad de modificaciones necesarias para poder eludir los sistemas de detección automática utilizados en las universidades (se puede ver una entrevista anónima con uno de estos estudiantes en [22]). En [14] se enlistan los niveles de dificultad de la detección según las modificaciones existentes en el texto plagiado, en orden de menor a mayor dificultad la lista muestra: “copia exacta del documento, copia de párrafos, copia de enunciados, copias con cambio de palabras, copias con reestructuración de las oraciones”.

2.2 Los rasgos observados en la detección automática de plagio

Existen algunas características observadas en la revisión de obras plagiadas que han dado pie a ser integradas en el análisis automático, motivando ciertas técnicas de detección de plagio. Estas características son enlistadas en [23], extendidas y revisadas por el mismo autor en [24]; las enunciamos (junto con unas pocas adicionales) mencionando brevemente cuál es el objetivo de su observación en el plagio:

1. Uso del vocabulario. El vocabulario usado en un nuevo documento debe de ser consistente con los documentos previamente presentados por el estudiante; la aparición de una gran cantidad de nuevo vocabulario no es común, en especial si éste es más avanzado que el de su nivel habitual.
2. Cambios de vocabulario. Estos cambios se refieren a los que suceden en el interior del documento; si párrafos o secciones del escrito contienen vocabularios distintos (provocados quizá por estilos o calidades diferentes), es probable que esto suceda porque han sido escritos por personas distintas, lo que abre la posibilidad de que alguna de esas porciones de texto sea plagiada.
3. Texto incoherente. Si a través del texto se presentan inconsistencias (como cambios de persona en la narración) puede deberse a que no haya sido escrito a conciencia o a que tiene insertos fragmentos escritos por algún otro autor.
4. Puntuación. Es improbable que dos autores distintos usen la puntuación de la misma manera.
5. Cantidad de texto común. Es habitual que documentos que abordan el mismo tema (inclusive aquellos que sólo son de temas relacionados) compartan cierta cantidad de texto, básicamente nombres y términos específicos del área; pero si se trata de documentos escritos de forma independiente, esta cantidad de texto similar o idéntico debería ser pequeña.
6. Errores ortográficos comunes. Una característica que evidencia el plagio es la existencia de errores ortográficos comunes, pues las fallas ortográficas (por ejemplo, los llamados “errores de dedo”) son descuidos del escritor y es inverosímil que también sean cometidos por otro autor, en especial en contextos similares.
7. Distribución de palabras. La distribución de las palabras se refiere a su frecuencia o habitualidad en determinado documento. Cada autor prefiere el uso de ciertos términos en lugar de otros; por lo que encontrar varias palabras que sean usadas con la misma frecuencia da pie a pensar que dichos textos están influenciados y que podría tratarse de un plagio.

8. Estructura sintáctica. Un indicativo del plagio es que los textos compartan las mismas estructuras sintácticas, referidas al estilo; se espera que la selección de la mayoría de las reglas sintácticas usadas por autores distintos sean diferentes.
9. Largas secuencias de texto común. Los textos escritos de forma independiente no deberían contener secuencias (de palabras o caracteres) comunes de gran longitud, incluso si abordan el mismo tema.
10. Orden de las similitudes textuales. Las coincidencias textuales que aparecen en dos documentos en el mismo orden, proporcionan mayor evidencia del posible plagio; claro que no hay que perder de vista que ciertas palabras, como los nombres de las secciones (introducción, cuerpo, resultado, conclusión, etc.), o expresiones que suelen usarse dentro de estas secciones (frases como: “el trabajo propuesto se...”, “este documento tiene la siguiente organización...”, “los resultados muestran...”, “se puede concluir después del trabajo desarrollado”, etc.) no proporcionan más argumentos al aparecer en cierto orden.
11. Preferencia por el uso de oraciones largas o cortas. Los autores suelen, por su estilo natural, tener alguna predilección por cierta longitud particular para sus oraciones. Las oraciones que están fuera del o los rangos de longitud habitual en el escrito constituyen secciones sospechosas que pueden ser investigadas con alguna otra característica.
12. Legibilidad del texto. Esto se refiere a la facilidad de lectura del texto. Algunos estilos se pueden leer de forma más ligera que otros. Se han planteado muchas medidas para cuantificar esta ligereza, algunos de ellos son los índices de *Gunning*, *Flesh* o *SMOG*, utilizados en ocasiones para evaluar si son textos adecuados para que puedan ser leídos por niños de alguna edad particular. Para más información de estos y otros índices puede consultarse [11].
13. Referencias incongruentes. La existencia de referencias a fuentes no enlistadas en la bibliografía o referencias numeradas que no se encuentra en el orden adecuado, sugiere que pudieron ser dejadas por descuido o ingenuidad al realizar un plagio.

14. Variaciones de formato. Si el documento tiene más de un formato (tipo y tamaño de letra, interlineado, justificación, etc.) a través del texto, puede deberse al copiado y pegado de secciones de otros trabajos.

La exposición de estos rasgos, que han servido de base para la creación o adaptación de los muchos métodos de detección automática del plagio, da un indicio del tipo de análisis que se suele hacer en esta tarea.

Posteriormente se presentará, en la sección 2.5, la forma particular en la que estas características se han tratado de capturar mediante ciertos modelos o técnicas particulares, pero antes es necesario introducir esas técnicas y algunos conceptos adicionales (secciones 2.3 y 2.4).

2.3 La problemática particular de la detección automática de plagio.

En la detección automática de plagio (DAP de aquí en adelante) es fundamental, para tener buena efectividad y eficiencia, considerar dos aspectos primordiales [5]:

- La base de datos en la cual se va buscar la fuente del plagio. Es importante tener una base de datos suficientemente extensa y particularmente dedicada o especializada (para asegurar que cubra con suficiente profundidad) al tema específico del que tratan los documentos que se van a analizar. Debido a que si no se tiene el texto de donde se ha robado el material, el plagio será difícilmente identificado y será aún más complicado obtener pruebas suficientes para reprender al plagiador.
- La estrategia (o método) de identificación. Obviamente aquí recae gran parte del desempeño del método; si bien puede ser relativamente sencillo conseguir los documentos que pudieron servir de fuentes, es definitivo que los *corpus* de las posibles fuentes por sí solos no sirven de nada sin un método apropiado de detección.

El número de fuentes con las que se trata ha hecho que se definan tareas específicas que solucionan distintas situaciones. En la figura 2.1 se presentan gráficamente esta diversidad de tareas distintas. Cada una de estas tareas tiene sus requerimientos y su particular objetivo [5]. Estas tareas son:

- i. Descubrimiento del plagio³. Aquí se determina el plagio en todo un *corpus*. Este tipo de análisis permite obtener todos los plagios que se hayan cometido dentro de un corpus particular; es ideal para determinar, por ejemplo, si se analizan los trabajos de un aula de clases, cuáles han copiado o plagiado algún texto de su o sus compañeros.
- ii. Búsqueda de plagio⁴. En esta tarea se tiene un único documento del cual se sospecha que ha cometido plagio y se somete a este análisis para encontrar la o las fuentes de donde ha obtenido el material plagiado.
- iii. Detección de plagio⁵. En la detección de plagio el análisis se hace sólo sobre un par de documentos: un documento sospechoso de contener o ser un plagio y un documento que se cree es la presunta fuente de este plagio. Siendo sólo necesario establecer si en ese par existe o no el plagio.
- iv. Detección intrínseca de plagio⁶. En esta tarea sólo se tiene un único documento, el sospechoso; se analiza todo el texto y los diferentes apartados, para determinar si contiene algún plagio. Generalmente en este análisis se obtiene la región o sección en donde se encuentra el texto plagiado (en caso de haberlo).

³ También se le conoce como detección corporal, pero fue nombrado en [3] como descubrimiento de plagio, lo cual se acerca más al trabajo que realiza.

⁴ El acuñamiento de esta tarea como “Búsqueda de plagio” fue dada en [30], pues originalmente en [5] es llamada detección multidimensional y en [31] es llamada detección extrínseca o externa de plagio.

⁵ Nombrada en [5] como detección apareada de plagio pero comúnmente conocida simplemente como detección de plagio.

⁶ Esta es una de las tareas particulares evaluadas en la competencia internacional de detección de plagio [31], en [5] es llamada detección singular (o “*single detection*”).

Tareas específicas automáticas del análisis del plagio









	Sospechosos	Fuentes
Descubrimiento		
Búsqueda		
Detección		
Detección intrínseca		

Figura 2.1. Se presentan las diversas tareas del análisis automático del plagio, estableciendo (el tamaño de) los conjuntos con los que trabaja como fuentes y como sospechosos.

La mayoría de los trabajos abordan la detección del plagio [5]. Una de las razones por la que se ha volcado el interés en esta tarea se debe a que la mayoría de las otras tareas se pueden solucionar de forma recursiva a través de la detección de plagio [27]. De manera que, la búsqueda de plagio se convierte en un problema de dimensionalidad 1:n (1 documento sospechoso y n fuentes) que se resuelve al aplicar la detección n veces, una para cada documento fuente. De forma similar el descubrimiento es un problema de n:n en donde se aplica n x n detecciones para analizar todos los posibles pares de documentos.

La detección intrínseca no puede modelarse a partir de la detección simple, o por lo menos, no se ha hecho con frecuencia; la generalidad de las soluciones intrínsecas utilizan las mismas herramientas que las empleadas en la tarea de atribución de autoría, pues se trata de una verificación de autoría a través del documento sospechoso.

Existen trabajos que no han realizado el descubrimiento de plagio de forma anidada, como en [27], donde se va procesando de forma simultánea todo el *corpus*.

En la búsqueda de plagio también se han planteado estrategias semi-anidadas, como lo hicieron casi todos los participantes de la competencia internacional de plagio [26], en donde la tarea se realiza en 2 etapas: una primera etapa que busca las k fuentes más probables del plagio y una segunda en donde se realiza k detecciones, una para cada uno de los pares conformado por una de las fuentes probables y el documento sospechoso (ejemplos de esta estrategia en la búsqueda de plagio son [28,29]). Esta división en dos etapas se debe principalmente a la gran variedad de técnicas de recuperación de información y a técnicas más refinadas, como [30,2], las cuales permiten encontrar muy eficazmente un subconjunto de documentos temáticamente relacionados con el documento sospechoso y, considerando que “el plagio usualmente ocurre entre documentos temáticamente relacionados” [31], esta selección no tendría por qué deteriorar la eficacia de la detección de plagio de la segunda etapa.

La DAP, a diferencia de la primera etapa de la búsqueda de plagio, requiere de un análisis más profundo [32,33], pues no sólo necesita determinar si los contenidos son de la misma temática; es necesario establecer qué tan relacionados están, e inclusive, si pudieron ser derivados uno del otro.

2.4 Trabajos previos en la detección automática de plagio.

En [5,34] se presenta una división muy elemental de los trabajos de acuerdo al enfoque del que parten para analizar la similitud entre dos textos. Se ha tomado esta división como base para establecer una jerarquía más completa del estado del arte en la DAP; esta jerarquía es presentada gráficamente en la Figura 2.2 y a lo largo de esta sección se explicará cada una de estas divisiones así como sus ventajas y desventajas. Esta jerarquía se centra en la variedad de representación de los textos usados por los diversos métodos de la DAP. Será hasta la sección 2.5 en donde se discutirá cómo es que los métodos deciden cuáles son los casos que consideran que son plagio.

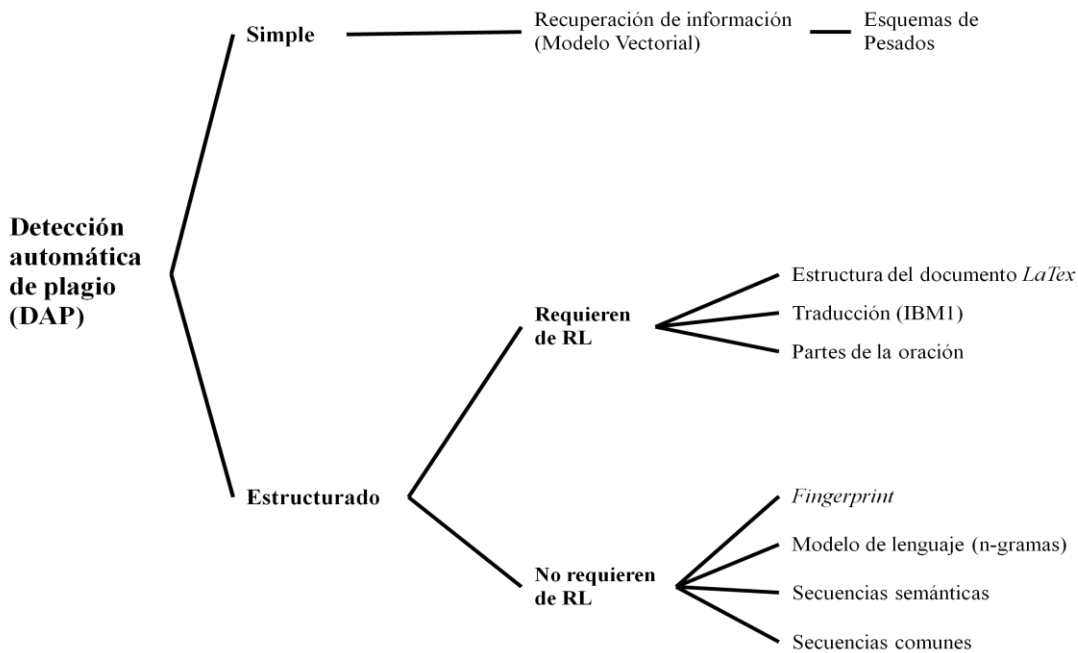


Figura 2.2 Jerarquía completa de las estrategias en la DAP.

La primera división de la jerarquía presentada proviene de [5,34], ahí se organizan los trabajos de acuerdo al enfoque del que parten para analizar la similitud entre dos textos. Los dos grandes enfoques básicos son el estructurado y el simple; el primero es llamado de esa manera pues tiene como principio que, al intentar capturar la estructura del lenguaje, obtendrá una mejor representación de las porciones que pueden estar plagiadas y así detectarlas más fácilmente. Por otro lado, el enfoque simple sólo se basa en obtener cierta información, a partir de las palabras que contiene el texto (sin importar el orden de éstas).

2.4.1 Enfoques Simples

Estos métodos son llamados simples pues tratan los documentos sencillamente como palabras sin estructura y usan el popular modelo vectorial comúnmente empleado en la recuperación de información (*IR*) [35].

Bajo este modelo se calcula la similitud de los textos, utilizando alguna medida que sólo involucra a las palabras que contienen; cuando esta herramienta se utiliza para la DAP (como en [36] y en el *baseline* de [22]) se tiene como primicia que los documentos con una alta calificación de similitud son plagios.

El modelo vectorial cuenta con algunos esquemas especiales de pesado; con éstos permite darle más importancia a algunas palabras del documento. El pesado de frecuencia del término, conocido como *tf*, permite aportar más valor a las palabras que aparezcan abundantemente, pues su repetición acentúa su importancia en el texto. El pesado *tf-idf* (*term frequency - inverse document frequency*) resalta la importancia de los términos “raros”, por lo que toma en cuenta el número de documentos que contienen esa palabra; entre menos documentos lo contengan más idiosincrático es este término [37].

Una de las medidas de similitud más populares para este modelo vectorial es la similitud coseno; proviene de una interpretación geométrica del modelo y básicamente mide qué tan parecido es un documento respecto al otro, utilizando el ángulo que existe entre sus representaciones vectoriales.

En esta cuantificación de la similitud no se presta atención a ninguna estructura, ni a la del documento ni a la de la oración, elimina todo rasgo existente en la coincidencia del orden en que se hayan escrito las palabras, esto es un terrible inconveniente para la DAP.

Otra forma de medir la similitud es el producto interno utilizado en [38], esta medida es la suma de los pesos de las palabras que tengan en común. Por tanto, los documentos son considerados más cercanos entre mayor número de palabras tengan en común y entre más valor se le asigne al peso de cada una de esas palabras. Una técnica con una filosofía similar que nació para atribución o verificación de autoría es la expuesta en [13]; en ella, a lo único que se le presta atención, es a la cantidad de términos que son empleados con la misma frecuencia en ambos documentos. Se espera que estas coincidencias estadísticas se deban al plagio.

Una estrategia popular en la DAP es seleccionar sólo un subconjunto de elementos en los cuales se calcula la similitud. Esta selección, en el modelo vectorial, típicamente se refiere a aquellas palabras que tengan frecuencia similares o estén dentro de un rango llamadas “clases de palabras” [39].

El modelo vectorial es sumamente efectivo para encontrar documentos relacionados con un tema de búsqueda que puede ser extraído de un texto particular (el documento sospechoso para la DAP), pero no resulta tan efectivo para encontrar aquellos documentos que han copiado o reutilizado texto [15]. El cambio del orden de las palabras

puede cambiar el sentido de una oración, y en el modelo vectorial éste es un hecho que no se toma en cuenta, siendo esto uno de los principales factores por lo que resulta tan poco efectivo en la DAP [40].

2.4.2 Enfoques estructurados dependientes de RL

Los enfoques estructurados mantienen, dentro del análisis, el hecho de que el lenguaje es naturalmente estructurado. A diferencia del enfoque simple, estos usan la estructura como un elemento adicional en la comparación de los textos; aparte de evaluar la similitud léxica (como el enfoque simple), centran su atención en comparar si las palabras ocurren en posiciones, funciones u órdenes similares, lo que podría implicar que sean escritos en formas similares.

Los métodos que engloba el enfoque estructurado se pueden dividir de acuerdo a su dependencia con recursos adicionales. Esta dependencia de Recursos del Lenguaje (RL) define el tipo de análisis que se hará con la estructura en el documento. Así los métodos que no dependen de los Recursos sólo pueden utilizar elementos estructurales muy simples, como el orden de las palabras; mientras aquellos que si usan RL podrán manejar la función sintáctica de las palabras u otro tipo de información.

Dentro de los que utilizan RL se encuentran aquellos que observan la estructura del documento entero [41] para analizar por separado sus secciones. Una gran desventaja de este enfoque es que los textos deben de estar escritos en *LaTeX* u otro lenguaje que presente etiquetas para indicar sus elementos, pues éstas permiten identificar durante el análisis las diversas partes de los documentos. Aparte del requerimiento de formato, otra desventaja es su relativa ceguera a los plagios que mezclen secciones del documento.

Por otro lado tenemos los métodos que están basados en la utilización de las herramientas de la traducción automática; básicamente modelan la DAP como un problema de evaluación de un texto traducido en el mismo idioma, en donde entre mejor sea la “traducción” más posibilidades hay de que exista un plagio [12,42]. Por último, existen algunas propuestas que provienen de la detección de plagio en software [4], en las que se utilizan herramientas automáticas del lenguaje, como los *parser* o analizadores que generan los árboles sintácticos de las oraciones. Estos árboles son empleados para comparar la

cantidad de estructuras sintácticas parecidas o iguales entre los documentos o para saber si las palabras son empleadas con las mismas funciones sintácticas.

El gran inconveniente de estos métodos es que requieren utilizar los RL adicionales [43], como la red léxica *WordNet* o grandes bases de datos bien balanceadas para el entrenamiento de los analizadores sintácticos o los analizadores de traducciones [42,44] y está subordinado a la eficacia de estas herramientas particulares (las cuales no son muy buenas para todos los idiomas).

2.4.3 Análisis estructural independiente de RL.

Los métodos que se aglutinan bajo la categoría del enfoque estructurado independiente de RL utilizan el orden y la adyacencia de las palabras para capturar (parcialmente) la estructura del lenguaje, aquí no es de interés conocer la función particular que desempeña cada palabra; en lugar de eso, capturan directamente el contexto en el que se encuentra. Esta estrategia resulta particularmente útil en la DAP, pues permite medir la cantidad de texto reutilizado e inclusive tener nociones acerca de la similitud de sus contextos.

Los modelos en esta categoría son muy simples, utilizan un análisis muy básico que consisten en la fragmentación del texto en trozos: *chunks*, *shingles* o *n-gramas*. Estos trozos tienen una cierta longitud llamada granularidad [38]. La granularidad se extiende por gran variedad de tamaños, desde un par de caracteres, unas cuantas palabras, o más grandes, como: una o varias oraciones, e inclusive, hasta párrafos o el documento entero. La selección de la granularidad tiene grandes repercusiones [39]; si se eligen trozos demasiado pequeños la probabilidad de que se repitan en otros textos será muy grande, sin importar que sean textos independientes (sin plagio); por otro lado elegir trozos muy grandes disminuye la posibilidad de que se encuentren en otro documento [38] y las pequeñas modificaciones o reescrituras como la omisión o cambio de alguna palabra, evitaría que las porciones plagiadas fueran detectadas [34].

Al conjunto de trozos generados a partir del documento son utilizados como sus características. Cuando se utilizan los *fingerprints*, estos trozos de texto son transformados a números a través de una función *hash* que garantiza con cierta probabilidad que el número

que se le asigna a un trozo particular no podrá ser asignado a otro trozo distinto a éste. El conjunto de estos números generados a partir de los trozos es la *fingerprint* del documento.

Los n-gramas son trozos de n palabras, el empleo de los n-gramas proviene de los modelos de lenguaje y su utilización en el reconocimiento del habla. Los *fingerprint* y los n-gramas son técnicas similares, con la diferencia de que los n-gramas no son transformados en números y que suelen ser de longitud menor, es decir, la n es generalmente pequeña. Los n-gramas, normalmente, se obtienen por trozos con traslape, en donde para tal caso en un texto de r palabras, se obtiene $r-n+1$ n-gramas.

Una estrategia aplicada con frecuencia a estos modelos es la selección de un subconjunto de los trozos, esto tiene, como principal objetivo, mejorar la eficiencia, pues al manejar solo una porción de fragmentos, las comparaciones serán menos costosas. Se espera que el subconjunto elegido pueda caracterizar suficientemente bien al documento para realizar en buen término la DAP y es por esto que existen múltiples propuestas para su elección. La estrategia de selección, que también es llamada resolución, obedece al hecho de que la mayoría de los plagios no tomarán el 100% del contenido de la fuente. Las diversas formas de selección [38,45] son:

- Selección por posición. Está basada en la posición que tienen los trozos dentro del documento; estas selecciones comprenden: la manipulación del traslape de los trozos, a través de su posición, para así extraer una menor cantidad. La elección aleatoria o de un cierto número de trozos en las primeras posiciones podría garantizar una distribución homogénea dentro del documento.
- Selección por frecuencia. Es una elección sustentada por la frecuencia de aparición de los trozos; entre estas estrategias se encuentran tomar los trozos más raros o escasos del documento o del *corpus* entero. Algunos métodos establecen la frecuencia por su prefijo o inicio.
- Selección por estructura. Esta intenta eludir los cambios de orden de los elementos plagiados; entre estas selecciones se encuentran: la estrategia que considera únicamente los trozos que contienen o que están inmediatamente después de un “ancla”. El “ancla” es un conjunto de sílabas, o incluso, hasta una

palabra que debe de ser lo suficientemente común para que exista al menos un trozo seleccionado en todos los documentos. Otra estrategia de selección de este tipo, aunque en extremo simple, es la que toma los trozos en la *K-esima* posición; seleccionan cada *k-esimo* enunciado o cada *k-esima* palabra de la oración o párrafo.

- Selección aleatoria. Se eligen los trozos que cumplen cierta regla elegida arbitrariamente, como quedarse sólo con los trozos que generará un valor de *hash* que sea divisible por un cierto número o que sea menor a algún otro número determinado aleatoriamente.

La elección de una resolución particular puede incrementar la eficiencia de la DAP. El inconveniente con todas estas estrategias es que ninguna puede garantizar que el plagio siempre sucederá dentro del subconjunto de trozos que ha separado como los representativos; incluso, ninguna de las estrategias está interesada en poder elegir aquellos que tengan más potencial de ser plagiadas. Más bien todas estas estrategias suponen, que si uno de los apartados del texto seleccionado es plagiado, es posible que existan otros, en el grupo de los no seleccionados, que también formen parte del plagio.

Las distintas resoluciones no son las únicas estrategias de selección o filtrado de los trozos; existen otras estrategias que buscan incrementar ya no la eficiencia, si no la eficacia de la DAP. En estas estrategias no se selección los trozos que son considerados para representar a los documentos, en su lugar se seleccionan los elementos que van a representar los trozos, con esto se intenta seguir reconociendo los casos de plagio aun cuando exista cierto cambio o reescritura del texto. La selección en estas estrategias está constituida por un subconjunto de palabras de cada trozo que se conservan como sus características relevantes. Esto permite a la DAP soportar el cambio u omisión de algunas palabras y por tanto seguir reconociendo los trozos reutilizados aun cuando no sean copias literales (plagio con cierta reescritura).

Dentro de estas ultimas estrategias de selección se encuentran los *k-skip-n-grams* [46], solución muy simple que realiza múltiples subconjuntos para una mismo trozo; genera todas las posibles selecciones que existen con n palabras representativas y hasta k palabras

omitidas, manteniendo el compromiso del mismo orden [42]. Eso implica que cada trozo de texto puede generar $k \times (n-1)$ *k-skip-n-grams*, haciendo explotar la memoria necesaria para almacenar los documentos. Otra solución que realiza una auténtica selección de los términos es la de las secuencias semánticas [40]. Estas secuencias son conformadas únicamente por los términos semánticamente relevantes. La discriminación de los términos relevantes, tomados como representativos, de los que no eran tomados, se realiza con un simple umbral de frecuencia, escogiendo las palabras poco comunes.

En todos los métodos en los que se utilizan ciertos trozos de texto de los documentos, el tamaño de los trozos es una variable que puede determinar en gran medida la calidad de la detección. Se ha visto que el uso de tamaños pequeños como los que suelen usarse en n-gramas ha resultado mejor [5], pero dentro de las posibles longitudes del n-gramas, es decir, el valor concreto de la n, existe aún la incógnita de cuál es el valor apropiado [47, 9, 32, 39]. Experimentalmente se ha encontrado una diversidad de posibles valores que parecen ser adecuados: en [32,43,48,49] los mejores valores de n fueron pequeños $n=2$ o 3 ; en [42,50] el mejor valor fue 4 , en [12] se halló que el mejor valor de n es el mínimo valor posible, n igual con 1 ; y en [33] se encontró que los mejores valores eran iguales con 12 . Esta gran disparidad de valores se debe, en buena medida, a que estos experimentos se realizaron en *corpus* de diversas naturalezas; en [33] se probaron los n-gramas en *corpus* de distintos idiomas y se encontró que los valores óptimos de la n para el caso del inglés era 12 ; para el alemán y el hindi, 10 , y para el español, 8 . Al utilizar n-gramas, cuya longitud se debe predefinir, se han tenido que realizar evaluaciones previas en secciones de *corpus* de entrenamiento para elegir un valor que aparentemente tiene una buena eficacia. Pero ha quedado patente, en los experimentos presentados en la literatura, que la dependencia de la naturaleza del *corpus* y del idioma del *corpus* hacen vulnerable al método ante cambios del entorno de evaluación.

Para hacer frente al problema de la elección del tamaño óptimo se ha propuesto usar tamaños no determinados o no homogéneos de los trozos, como en [45], donde la longitud de los trozos es determinada por el uso de puntos de corte dinámicos. Estos puntos de corte separan los trozos utilizando la función *hash* para decidir cuándo se realiza un punto de corte. En el momento en que el valor *hash* es divisible entre un cierto k es cuando

se realiza el corte (esta condición es conocida como: $0 \bmod k$). Otra propuesta para el problema de la elección del tamaño óptimo ha sido usar más de un sólo valor predefinido de longitud de los trozos (generan varios n-gramas con distintos valores de n) [9].

Las propuestas anteriores sólo evaden el problema de la selección de la longitud óptima; $0 \bmod k$ no propone una longitud óptima, sólo deja la elección a un proceso aleatorio y la selección de varias longitudes se vuelve un problema doble, pues no todo par de valores es una buena elección y no se puede garantizar que los conjuntos de detecciones realizadas, con cada uno de los valores de n, sean complementarios, y por tanto que la unión ayude a aumentar la eficacia de la detección. Una mejor aproximación para la elección de trozos de longitud variable es el tratamiento como cadenas o *strings*.

La última estrategia de análisis, sin dependencia de RL, que presentaremos proviene de la bioinformática, en particular de la búsqueda de elementos comunes entre secuencias genéticas. En esta tarea, primero, se utilizó la búsqueda de la secuencia común más larga (Longest Common Subsequence, LCS) [42,51]; después se desarrollaron algoritmos que eran menos costosos que los de la búsqueda de la LCS; solucionaban la búsqueda de los elementos comunes de manera local. Dentro de estas nuevas propuestas estaban la YAP [52], y sus dos siguientes versiones la YAP2 y la YAP3, esta última también llamada GST (*Greedy String Tiling*), en ella se encuentran coincidencias exactas (*tiles* o secuencias comunes) entre dos cadenas.

La GST se ha usado en la DAP [12] para encontrar las secuencias de palabras que tienen en común el documento sospechoso y la fuente. Con estos *tiles* es posible reconocer en buena medida al texto reutilizado. La GST emplea un umbral de longitud para evitar confundirse entre secuencias comunes que no sean propias de la reutilización, como son términos comunes debido a la temática, lo que implica que en este método se espera que las porciones de texto reutilizado sean superiores a ese umbral.

Estas secuencias comunes, o *tiles*, también han sido llamadas “*true matches*” [28]; la diferencia de nombres se debe a la forma en la que se extraen. Existen diversos algoritmos con diferentes exigencias de complejidad en tiempo y espacio [28,53]; para uniformar a estas secuencias, a partir de aquí llamaremos simplemente secuencias comunes

a estas secuencias de máximo tamaño que se encuentran de forma idéntica en ambos documentos (fuente y sospechoso).

Las secuencias comunes hacen frente a las dos grandes dificultades de las técnicas del análisis de estructura bajo: la granularidad y la resolución. La primera es autoajustada a cada una de las secuencias comunes, según la coincidencia lo exija; y la resolución se realiza al estar sólo capturando las secuencias que se comparten en ambos documentos y automáticamente se descartan todas aquellas que no estén compartidas.

2.5 Los rasgos observados en los diferentes métodos de DAP

Las características del plagio planteadas en la sección 2.2 y resumidas en el tabla 2.1 son rasgos que se han intentado capturar en diversos métodos de detección. Estas características han sido centro de atención en algunos métodos como se puede ver en la Figura 2.4, pues son aspectos que develan el plagio.

1. Uso del vocabulario.	Observa el vocabulario habitual en busca de palabras inusuales para ese autor.
2. Cambios de vocabulario.	Consistencias léxicas y de estilo a través del documento.
3. Texto incoherente.	Inconsistencias (como cambios de persona o tiempo en la narración).
4. Puntuación.	Estilo ortográfico similar
5. Cantidad de texto común.	Texto existente en ambos documentos
6. Errores ortográficos comunes.	Ejemplo de estos errores ortográficos son los “errores de dedo” u otros descuidos.
7. Distribución de palabras.	La frecuencia o habitualidad de las palabras en el documento.
8. Estructura sintáctica.	Oraciones construidas por estructuras con los mismos patrones sintácticos.
9. Largas secuencias de texto común.	La extensión de las coocurrencias textuales determina potenciales plagios.
10. Orden de las similitudes textuales.	La secuencia de las coincidencias textuales individuales es evidencia importante.

11. Preferencia por el uso de oraciones largas o cortas.	Por estilo de forma natural se suelen preferir oraciones con cierta extensión.
12. Legibilidad del texto.	Esto es la facilidad para su lectura. Cuan ligero y asequible es leer el texto; medido por índices particulares.

Tabla 2.1. Resumen de rasgos examinados en diversos métodos de la DAP.

La Figura 2.4 muestra la forma en la que los diversos métodos utilizan las características de la tabla 2.1. Algunos de los métodos prestan atención a más de una de estas características y en otras ocasiones el método no fue, en principio, diseñado para observar la afectación de ciertos rasgos, sin embargo, por su naturaleza es sensible a éstos.

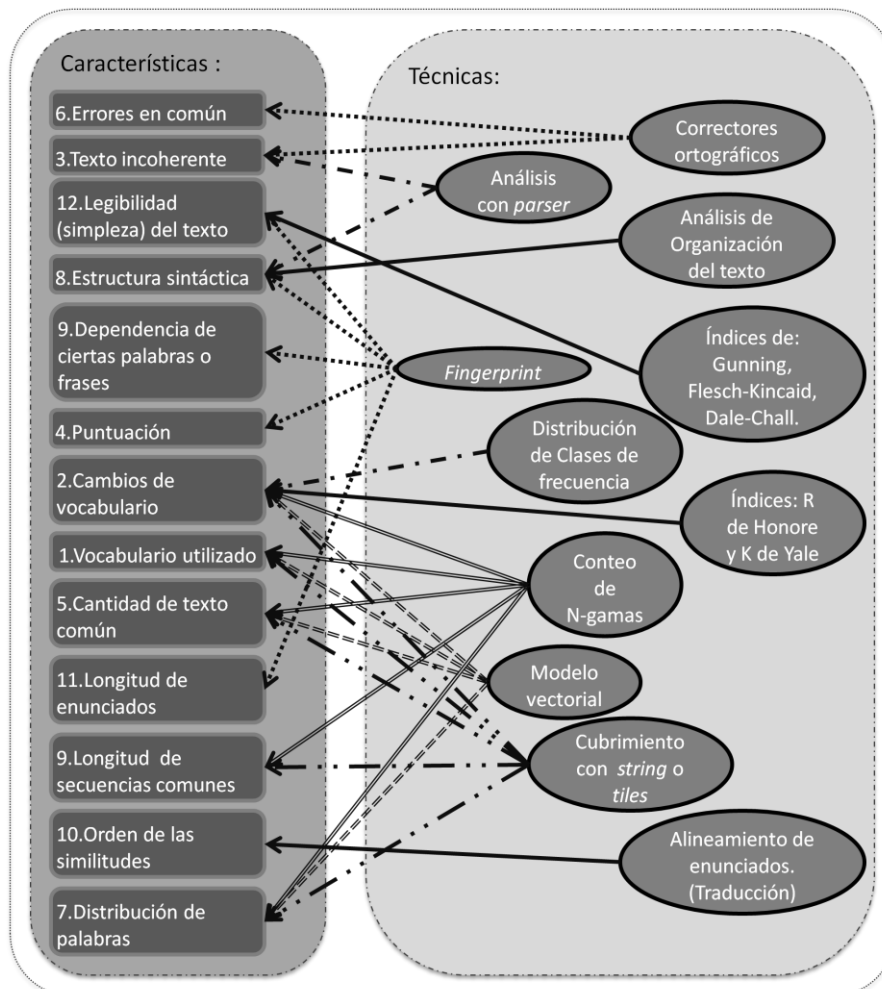


Figura 2.4 Se esquematiza la existencia de múltiples técnicas que responden a muchas de las características especificadas en la sección 2.1. Las técnicas están simbolizadas por los óvalos y las características por los rectángulos, las flechas unen los aspectos que observa cada técnica.

Las características 1, 2, 5, 6 y 7 son las principales motivaciones de la utilización del modelo vectorial en la detección del plagio. El modelo vectorial es muy popular en otras áreas del procesamiento del lenguaje natural, permite capturar de manera aislada las palabras que ocurren en el documento para poder extraer algunas conclusiones a partir de la ausencia o uso de las palabras en los documentos.

Además de las características que motivaron el uso del modelo vectorial (1, 2, 5,6 y 7), la técnica conocida como n-gramas (que es la representación de los documentos a través de secuencias de n palabras consecutivas) permitió tomar ventaja de las características 5 y 9. Sin embargo, la mejor forma de utilizar la característica 9 es mediante el uso de *tiles* (secuencias consecutivas, pero en lugar de ser de tamaño fijo n, pueden tener cualquier tamaño).

Las características 3 y 6 pueden ser aprovechadas con herramientas de análisis y ayuda para el lenguaje natural; los correctores automáticos de ortografía y los *parsers* (analizadores sintácticos), en general, son características poco utilizadas pues no parecen dar los mejores resultados por sí mismos.

Las características 8, 11 y 12 pueden ser monitoreadas en los métodos estadísticos, principalmente a través del análisis de estilo.

Finalmente, las características 13 y 14 no son características estándares en la detección automática; la razón por la que habitualmente no son empleadas es debido a que su origen está en los descuidos del plagiador, descuidos que con las nuevas herramientas de ediciones de texto son menos frecuentes.

2.6 Decisión de plagio en la DAP

Todas las estrategias expuestas en 2.3 son estrategias para modelar la reutilización del texto, pero aún queda pendiente discutir cómo es que, a partir de alguno de estos análisis del texto posiblemente reutilizado, es que se decide cuál documento es plagiado y cuál no.

La idea patente en prácticamente todas las estrategias es obtener el texto que es, en apariencia, reutilizado mediante la intersección de los documentos; por tanto, el número de

unidades, ya sean palabras, oraciones o trozos que se encuentren en ambos documentos, se le llamada intersección.

Con la intersección se obtiene una medida de qué tan aparentemente copiado está el documento. Existen varias propuestas de estas medidas [35], la diferencia primordial es la normalización en la medida. Algunas medidas sólo toman en cuenta el documento sospechoso, mientras otras consideran el tamaño de ambos documentos. El coeficiente Dice y Jaccard (usados en [43]) está normalizado sobre la suma del tamaño de los dos documentos, mientras que *Resemblance* (en [54]) está normalizada con el tamaño del documento más pequeño. Por otro lado *Containment* (empleado en [2, 12, 32, 54]) sólo toma en cuenta al documento de interés, es decir, el sospechoso.

Tomando como nomenclatura D^s como documento sospechoso y D^f , documento fuente. Tenemos $|D^x|$ como el tamaño o número de palabras contenidas en el documento D^x ; $trz(D^x)$ es el conjunto de los trozos, ya sean *chunks*, *shingles*, *n-gramas* o *fingerprint*, que se generan a partir del documento D^x . Los diferentes coeficientes son:

$$Jaccard = \frac{|trz(D^s) \cap trz(D^f)|}{|trz(D^s) \cup trz(D^f)|} \quad (2.1)$$

$$Dice = 2 \frac{|trz(D^s) \cap trz(D^f)|}{|trz(D^s)| + |trz(D^f)|} \quad (2.2)$$

$$Resemblance = \frac{|trz(D^s) \cap trz(D^f)|}{\min(|trz(D^s)|, |trz(D^f)|)} \quad (2.3)$$

$$containment = \frac{|trz(D^s) \cap trz(D^f)|}{|trz(D^s)|} \quad (2.4)$$

La medida más popular para la DAP es *Containment* (2.4), ya que, en la mayoría de los casos, no resulta tan importante la longitud de la fuente.

Todas las medidas anteriores dan una valoración del plagio; tienen como máximo valor 1 y como mínimo valor 0 [35,55]. Se espera que cuando un documento tiene un valor

de 0 es un documento no plagiado u original y cuando tiene un valor de 1 es un documento plagiado. El gran problema es determinar qué pasa cuando la medida está entre 1 y 0. La inmensa mayoría de las propuestas previas utilizan un umbral sobre estas medidas. Para establecer el valor de este umbral se usan dos posibles estrategias:

- i. Umbral de valor fijo [10,18]. Se utiliza un valor arbitrario o previamente determinado por un experto, como puede ser por ejemplo: todos los documentos que tengan más del 50 % de su contenido como reutilizado serán considerados plagio.
- ii. Umbral determinado para el *corpus* [32]. Se selecciona el valor al evaluar una sección del *corpus* de documentos para poder determinar el punto óptimo de funcionamiento.

Existe un pequeño grupo de propuestas que, en lugar de emplear un umbral, ha utilizado técnicas de aprendizaje automático [12,51]; esto ha permitido mejorar el desempeño de la DAP y realizar una clasificación de los casos de plagio y no plagio que no necesariamente recurrían a una separación lineal. También abre la posibilidad de crear métodos de análisis más complejos que puedan combinar información de manera sofisticada [23], pero las propuestas existentes sólo han hecho métodos que únicamente combinan los resultados de varias técnicas previas de DAP. Estas combinaciones toman a cada método como un atributo particular [51]. Son propuestas interesantes pero que no explotan todo el potencial de las capacidades del aprendizaje automático para poder inferir cierta característica de los objetos (aquí sería la condición de plagio o no plagio) a partir de varios atributos que lo describan. En realidad, las propuestas existentes son combinación más cercanas al *Stacking* [56] de las técnicas previamente propuestas que a un método de DAP basado en el aprendizaje automático.

3 El método propuesto

3.1 Introducción al método

El método propuesto constituye un nuevo enfoque para la DAP. En él se establece una manera de caracterizar la similitud entre los textos mediante atributos que describen la semejanza. Estos atributos no sólo capturan qué tanto se asemejan los documentos sino, también, cómo es su similitud. Una descripción más amplia de la similitud permitirá a los algoritmos de aprendizaje automático poder distinguir de mejor manera los casos que son plagio de aquellos que no lo son, y hasta determinar el grado o tipo de plagio.

Las innovaciones del método propuesto se pueden englobar en 3 ejes principales:

- Una forma de estructurar la evidencia disponible para la DAP, conformando una representación que lleva a la DAP a un problema de aprendizaje automático.

El uso de la representación implica cambiar la visión del problema; de la tradicional cuantificación del supuesto material plagiado, a su cualificación⁷.

- Un enfoque mixto, empleado en la generación de la representación de la similitud. El enfoque mixto aprovecha las ventajas de los principales enfoques de la DAP, el simple y el estructurado (secciones 2.4.1 y 2.4.2).

El enfoque simple normalmente sólo utiliza los pesos que se les asigna a las palabras según su importancia. De este enfoque se ha tomado la idea de poder utilizar pesos para enfatizar los elementos que sirven como evidencia importante en la DAP.

Por otro lado, utilizamos el enfoque estructurado al usar las secuencias comunes para capturar las posibles porciones del plagio. La mezcla de los enfoques simple y estructurado se encuentra en la forma del pesado que evalúa todo el fragmento estructurado y lo pesa como una unidad y como un conjunto de palabras, creando un mutuo compromiso de estos dos aspectos.

- Una nueva forma de obtener las secuencias de texto reutilizado, que generalizan las secuencias de texto consideradas como evidencia de plagio, a partir de flexibilizar las condiciones de similitud. Estas porciones de texto llamadas por nosotros “cadenas”, no cumplen la definición estricta de secuencia de las propuestas anteriores. Para su obtención se desarrolló una nueva valoración de la reutilización de cada palabra llamada Índice de Reescritura.

Como ya se ha mencionado en las secciones anteriores, todos los documentos tienen cierto texto en común debido a palabras o construcciones propias del idioma⁸. También se debe tener en cuenta que, en especial, los documentos sospechosos y sus fuentes están temáticamente relacionados, lo que implica una buena cantidad de términos temáticos compartidos. Estas coincidencias normales no implican plagio, pero, para un sistema automático, es muy difícil reconocer cuáles son las coincidencias normales y cuáles

⁷ Cabe aclarar que la cualificación de los casos de plagio, debido a las posibilidades del procesamiento automático de textos, se realiza mediante varias cuantificaciones que miden ciertas características cualitativas. Las llamamos cualificaciones, pues las medidas se refieren a la existencia y forma de ciertas características.

⁸ Las construcciones propias del idioma son frases como: “la prueba es la siguiente” [15] o las enlistadas en la característica número 10 de la sección número 2, la existencia de este tipo de frases en el documento sospechoso y en la fuente no aportan mucha evidencia pues son frases comunes.

son verdaderas evidencias del plagio. Las porciones de texto que los sistemas automáticos consideran como una reutilización, cuando en realidad son coincidencias normales, son llamados falsos positivos; y son frases que no aportan información para determinar el plagio [10].

La forma de estructurar la evidencia y de pesarla, propuestas en este trabajo, tienen como objetivo dar la suficiente información a los algoritmos de aprendizaje automático para identificar los patrones de las porciones de texto que conforman los falsos positivos.

El caso opuesto al fenómeno de los falsos positivos son los falsos negativos. Estos ocurren cuando una porción de texto ha sido plagiada de la fuente y modificada severamente, haciéndola pasar para los sistemas automáticos por un fragmento original [5]. Las modificaciones realizadas al texto plagiado son operaciones de reescritura, como el cambio de palabras por sinónimos o el reordenamiento de la oración. La generalización del texto reutilizado, mediante nuestra propuesta de cadenas, fue diseñada para bajar las tasas de los falsos negativos, y en consecuencia lograr una mejor detección del plagio.

3.1.1 Secuencias comunes

En esta sección introduciremos las secuencias comunes, una herramienta que muchos trabajos previos utilizan en la DAP. Hacemos mención de las secuencias comunes pues nos ayudarán a explicar de forma sencilla algunas propuestas de la tesis, así como a evaluarlas de una mejor manera, tomando en cuenta la forma en la que se realiza habitualmente la recolección de la evidencia del plagio. Antes de empezar a explicar que son las secuencias comunes debemos advertir al lector que esta forma de recolectar la evidencia no es parte del método propuesto y que estas secuencias comunes serán reemplazadas por las cadenas obtenidas por el Índice de Reescritura en el método propuesto. Las cadenas del Índice de Reescritura y el Índice de Reescritura son conceptos fundamentales de la tesis y serán explicados a detalle posteriormente, en las secciones 3.4.1, 3.4.2 y 3.4.3.

Las secuencias comunes son las porciones de texto idénticas entre ambos documentos, lo que nos permite obtener la evidencia tangible para detectar el plagio. Mediante estas secuencias se obtienen todas las coincidencias, tal y como aparecen en los documentos, identificando todo el texto reutilizado en forma exacta.

Sabemos que no toda secuencia común corresponde a un plagio, ya que existen los falsos positivos. Para lidiar con este problema se propone usar un conjunto de atributos que describan diversas características de las secuencias comunes para poder calificarlas y con ello determinar si son o no parte de un documento plagiado.

Antes de describir los atributos propuestos, definimos las secuencias comunes que es la evidencia que describirán los atributos; tomando a D^s y D^f como documentos, sospechoso y fuente, respectivamente, y suponemos que cada documento es una secuencia de palabras donde w_i^s y w_i^f son las i -ésimas palabras de D^s y D^f respectivamente, luego:

Definición 1. La secuencia de palabras $s^s: \langle w_i^s, w_{i+1}^s, \dots, w_{j-1}^s, w_j^s \rangle$ contenida en D^s es una secuencia común entre D^s y D^f si y sólo si existe al menos una secuencia $s^f: \langle w_{i+x}^f, w_{i+x+1}^f, \dots, w_{j+x-1}^f, w_{j+x}^f \rangle$ en D^f , tal que:

$$\forall_{i \leq k \leq j} w_k^s = w_{k+x}^f \quad (3.1)$$

$$w_{i-1}^s \neq w_{i+x-1}^f \quad (3.2)$$

$$w_{j+1}^s \neq w_{j+x+1}^f \quad (3.3)$$

La primera condición 3.1 asegura que cada elemento de la secuencia s^s tiene un elemento igual en el mismo orden⁹ en la secuencia s^f ; por tanto las secuencias son iguales. Las condiciones 3.2 y 3.3 aseguran que la secuencia común es de longitud máxima y que ninguna subsecuente sea también una secuencia común.

El conjunto de todas las secuencias comunes entre D^s y D^f es denotado Ψ , y conforma la principal evidencia con la cual se debe determinar si existe plagio. Para poder discriminar los casos de plagio de los que no lo son, se propone caracterizar Ψ mediante un

⁹ El mismo orden se refiere a que cada elemento igual tiene el mismo subíndice interno, el k en la ecuación 3.1, en la secuencia.

conjunto de atributos dividido en 2 principales tipos: los llamados atributos de distinción y los atributos de fragmentación.

3.2 Atributos de fragmentación

Estos atributos conforman una nueva manera de estructurar la evidencia (constituida por todas las secuencias comunes, Ψ) para la DAP. Normalmente, el manejo de la evidencia es muy pobre, ésta únicamente se evalúa considerando la cantidad de texto del documento sospechoso que se ha considerado como potencial evidencia del plagio, y por tanto, texto incluido en Ψ . Esta simple consideración existente en las técnicas tradicionales, implica que lo único importante para determinar el plagio es la cantidad de palabras que contenga todo el conjunto de evidencias Ψ . Habitualmente, como se hace en diferentes medidas de similitud (sección 2.4), el monto de palabras en Ψ se contrasta con el tamaño del documento sospechoso para medir el porcentaje de palabras del sospechoso que son encontradas como potencial evidencia.

El uso de las medidas de similitud colapsa¹⁰ toda la información inmersa en la constitución de Ψ a una sola cantidad; estas medidas sólo proporcionan el porcentaje de traslape existente entre los documentos, pero para la DAP la verdadera meta es saber la relación entre los documentos [33], discriminar si el texto común es parte de un plagio y (en ocasiones) especificar qué tipo de plagio es. Para evitar esta pérdida de información en las técnicas tradicionales se proponen usar los diversos atributos.

Los atributos de fragmentación propuestos se centran en dos aspectos importantes para el plagio: qué tan fragmentado está Ψ (es decir, cuán particionado es el texto reutilizado) y qué tamaño tienen esos fragmentos. Resulta evidente que un fragmento de texto de gran tamaño es un elemento más concluyente de un plagio que un fragmento

¹⁰El colapso de toda la información es la pérdida de información de Ψ debida a la representación de este conjunto de secuencias con un único dato que refleja la cantidad de palabras que contiene con respecto al documento entero. Esta simplificación pasa por alto valiosa información implícita en la composición del conjunto de las secuencias comunes Ψ .

pequeño, e inclusive, más concluyente que varios fragmentos pequeños; un fragmento grande, entre más grande sea, más claro es que se trata de una porción de texto que tuvo que ser copiada para que coincidiera de forma literal en ambos documentos; mientras que, la existencia de un mayor número de fragmentos pequeños puede explicarse por una cercanía temática más estrecha. Estas primicias, que podrían parecer obvias, son característica normalmente ignoradas. En el trabajo aquí presentado el tamaño de los fragmentos reutilizados es aprovechado para caracterizar los casos de plagio, conformando uno de los grupos de atributos del método propuesto; los atributos de fragmentación enfatizan las diferencias de tamaño en las diversas secuencias comunes de Ψ .

La forma en la que se solucionó la representación para poder expresar las características de la constitución de Ψ , es mediante una valoración separada de las secuencias comunes. Dicha separación se realizó mediante el tamaño de cada secuencia común. Se crean subconjuntos de secuencias comunes ψ_i ; cada subconjunto ψ_i solo contiene secuencias comunes (s) de tamaño i (ecuación 3.4).

$$\psi_i = \{s: s \in \Psi \wedge |s| = i\} \quad (3.4)$$

El conjunto Ψ puede contener gran cantidad de secuencias comunes y éstas pueden tener tamaños muy diversos, cada posible tamaño genera un nuevo subconjunto, por lo que, para no manejar una infinidad de subconjuntos, definimos un subconjunto Ψ_m ¹¹ que contiene todas las secuencias comunes tamaño mayor o igual a un umbral m . Este subconjunto Ψ_m aglutina a todos los subconjuntos ψ_i con un subíndice que va de m a la menor de las longitudes de los documentos fuente o sospechoso ($\min(|D^s|, |D^f|)$).

$$\Psi_m = \{s: s \in \Psi \wedge |s| \geq m\} \quad (3.5)$$

Por tanto Ψ se encuentra dividido en múltiples subconjuntos:

$$\Psi = \{\psi_1, \psi_2, \psi_3, \dots, \psi_{m-1}, \Psi_m\} \quad (3.6)$$

Y Ψ_m está compuesto por todos los ψ_i de la secuencias mayores o iguales que m ($i \geq m$).

¹¹ Para una justificación de la existencia del subconjunto Ψ_m véase más adelante.

$$\Psi_m = \{\psi_m, \psi_{m+1}, \psi_{m+2}, \psi_{m+3}, \dots, \psi_{\min(|D^s|, |D^f|)}\} \quad (3.7)$$

Los conjuntos ψ_i ahora son subconjuntos especializados de Ψ , el conjunto de las evidencias, por tanto, se ha organizado la evidencia en los subconjuntos ψ_i . Debido a que el criterio de separación de los subconjuntos es la longitud de las secuencias comunes, entonces la organización de la evidencia responde a la longitud de las secuencias comunes. Ahora es posible contabilizar las palabras en cada subconjunto y obtener una serie de atributos que organizan la evidencia tomando en cuenta el tamaño de las coincidencias del texto; estos atributos son los llamados atributos de fragmentación.

Con cada uno de los subconjuntos de secuencias comunes ψ_i se genera un atributo que está especializado en las secuencias comunes de una longitud particular en ψ_i . El conjunto Ψ es mapeado a una lista de atributos, donde cada uno de ellos está dedicado a un tamaño de secuencia común particular; (en otras palabras, sólo considera a un ψ_i específico); si f_i^{frag} es un atributo de fragmentación para las secuencias comunes de tamaño i entonces:

$$\Psi \rightarrow \langle f_1^{frag}, f_2^{frag}, \dots, f_m^{frag} \rangle \quad (3.8)$$

Cada atributo f_i^{frag} (3.9) es calculado como la suma de los tamaños de las secuencias comunes (s) en Ψ de tamaño i .

$$f_i^{frag} = \sum_{\{s_j: s_j \in \Psi \wedge |s_j|=i\}} |s_j| \quad (3.9)$$

O lo que es lo mismo en la notación de subconjuntos (3.10) al número de secuencias contenidas en ψ_i multiplicado por su tamaño i .

$$f_i^{frag} = i \cdot |\psi_i| \quad (3.10)$$

Las secuencias comunes grandes son poco abundantes y, entre más grandes, menos abundantes son; por lo que no tiene sentido dedicar un atributo conjuntos vacíos o con muy pocas secuencias comunes. Se aplico un límite al tamaño máximo que contempla

la representación pues de lo contrario se tendría en la representación una gran cantidad de atributos en cero, lo cual no es deseable para el aprendizaje automático¹².

Al definir el límite, el último atributo con el subíndice m aglutina todas las secuencias mayores o iguales a m . En el cálculo del f_m^{frg} se considera la longitud de cada una de las secuencias que engloba, con el objeto de dar la importancia apropiada a las secuencias grandes, por lo que la forma de la ecuación para calcularlo es muy parecida al resto de los atributos, con la excepción de los tamaños de las secuencias consideradas (ecuación 3.11).

$$f_m^{frg} = \sum_{\{s_j: s_j \in \Psi \wedge |s_j| \geq m\}} |s_j| \quad (3.11)$$

O en forma similar (3.12) considerando la notación de los subconjuntos.

$$f_m^{frg} = m \cdot |\psi_m| + (m + 1) \cdot |\psi_{m+1}| + (m + 2) \cdot |\psi_{m+2}| + \dots \quad (3.12)$$

3.3 Atributos de distinción

La segunda novedad en el método propuesto es la utilización de las ideas de la IR en la valoración de las secuencias comunes. En la IR se usan pesos para evaluar la importancia que tienen los términos en un documento para describirlo y/o discriminarlo. Aquí utilizamos esa idea aplicando un pesado a cada secuencia común; este pesado permite estimar la relevancia para el plagio, tomando en cuenta información adicional a la existencia común de la secuencia en ambos documentos.

En algunas técnicas previamente propuestas y reportadas en el estado del arte, se realiza una selección de las secuencias comunes que se consideran al hacer el análisis del

¹² . Para seleccionar este límite de tamaño máximo de la secuencia (denotado con m para ser congruente con el límite de los subconjuntos y con el mapeo de la representación 3.7) empleamos la ganancia de información (G.I.), una medida que nos permite observar qué tan bien separa las clases los diferentes atributos. Auxiliándonos de la G.I. [62] podemos determinar a partir de qué tamaño las secuencias comunes se deberían dejar de considerar por separado.

plagio. La idea de seleccionar sólo algunas secuencias comunes se conoce como la manipulación de la resolución, y puede ser vista como una valoración o pesado binario de las secuencias donde, o son tomadas en cuenta (peso de 1), o no son tomadas en cuenta (peso de 0). En lugar de la tajante selección binaria que descarta a todas aquellas que no cubran cierto requisito de selección; en el método propuesto se formula una solución con mayor riqueza en la valoración de las secuencias comunes.

La idea básica es que se puede ponderar a las secuencias comunes con base en las palabras que las constituyen. Se parte de una suposición muy frecuente en el procesamiento automático de textos: la frecuencia de las palabras en los textos nos pueden dotar de información acerca de la importancia de esas palabras para el mismo texto. En el caso particular del plagio es importante saber qué tan único es lo que parece ser copiado, ya que, si en realidad se trata de una oración o término frecuente, la probabilidad de que esas palabras hayan sido copiadas es menor. Por tanto, las palabras muy frecuentes son menos distintivas y por tanto no son relevantes para el plagio [38].

En el método propuesto, debido a que las secuencias comunes son palabras contiguas que forman expresiones, se utiliza el compromiso de todas las frecuencias de los términos¹³ que forman a la secuencia común s ; por tanto, se observa cuántas veces ocurre cada una de las palabras, sin importar si están juntas y en el orden adecuado. Y por otro lado, también, se usa la frecuencia de la secuencia completa, es decir, cuántas veces aparece la secuencia común tal como es, con todas sus palabras en el orden apropiado.

El pesado de distinción (3.13) propuesto aplicado a una secuencia (s_i) de tamaño $|s_i|$ es:

$$distinción(s_i) = \frac{1}{e^{freq(s_i, D^s) - 1}} \times \prod_{k=1}^{|s_i|} \frac{2}{freq(w_k^{s_i}, D^s \cup D^f)} \quad (3.13)$$

¹³ El compromiso de las palabras de la secuencia común es modelado como la multiplicación de la frecuencia inversa; la multiplicación de las frecuencias de las palabras se ha utilizado en algunas otras tareas, como la evaluación de las traducciones que hace el modelo M1 de IBM.

Donde $freq(s_i, D^s)$ indica las ocurrencias de la secuencia común s_i en el documento D^s , y $freq(w_k^{s_i}, D)$ indica las veces que aparece en el documento D , la palabra k , perteneciente a la secuencia s_i , $w_k^{s_i}$. Para el caso de la última expresión, D en el pesado de distinción es $D^s \cup D^f$ que se puede ver como el documento resultante de la concatenación de ambos documentos D^s y D^f o como la suma de las frecuencia de la palabra $w_k^{s_i}$ en cada documento.

Esta medida de relevancia, el pesado de distinción (3.13), tiene dos componentes: el primero¹⁴ evalúa qué tan probable es la expresión de la secuencia completa en el dominio sospechoso, castigando severamente el re-uso de la expresión de forma literal, pues se espera intuitivamente que el material plagiado no sea usado reiteradamente. En el denominador se le resta la unidad al exponente para normalizar la expresión, recordando que la función de distinción sólo se aplica a las secuencias comunes y, por tanto, la frecuencia en el documento sospechoso $freq(s_i, D^s)$ es por lo menos igual a uno, por tanto, este primer término tendrá siempre valores en el rango $(0,1]$.

El segundo componente de 3.13, la multiplicación, intenta capturar qué tan común o probable es cada término (por separado) de la secuencia. Se toma en cuenta la ocurrencia del término tanto en D^s como en D^f porque éste es el universo temático con el que se cuenta¹⁵. Se calcula la multiplicación de la frecuencia inversa de cada término.

La medida de distinción está normalizada; los posibles valores para cada secuencia están en el rango $(0,1]$ en donde el máximo, 1, sólo se obtiene cuando la secuencia común aparece una sola vez en los documentos y sus términos conformantes sólo fueron usados en ella. Este caso es el de la valoración máxima de distinción, pues la frase solo aparece una única ocasión al igual que los términos con los cuales está conformada.

¹⁴ En la forma de la expresión matemática de este primer componente se encuentra el inverso de la función exponencial; el uso de esta función matemática para la evaluación de posiciones de texto es también usada en [20], aunque en una forma muy diferente. En ambos casos, el propuesto y el del trabajo presentado en [20], la primicia es castigar con rápido crecimiento la existencia de copias idénticas.

¹⁵ En [44] también se realizan cálculos de la frecuencia de las palabras en la unión de ambos documentos, el sospechoso y la fuente.

Mediante el uso de la medida de distinción 3.13 se calcula un conjunto de atributos de distinción. Al igual que los atributos de fragmentación, cada uno de éstos está dedicado a un tamaño de secuencia particular. El cálculo de un atributo de distinción f_i^{dist} (3.14) para las secuencias de tamaño i es:

$$f_i^{dist} = \sum_{\{s_j: s_j \in \Psi \wedge |s_j|=i\}} distinción(s_j) \quad (3.14)$$

De igual modo se tiene un atributo f_m^{frg} (3.15) que aglutina todas las secuencias mayores o iguales a m .

$$f_m^{dist} = \sum_{\{s_j: s_j \in \Psi \wedge |s_j| \geq m\}} distinción(s_j) \quad (3.15)$$

Por tanto, el conjunto de las secuencias comunes Ψ se mapea al conjunto de atributos (3.16) de distinción y fragmentación que constituyen la representación de los posibles casos de plagio.

$$\Psi \rightarrow \langle f_1^{dist}, f_2^{dist}, \dots, f_m^{dist}, f_1^{frg}, f_2^{frg}, \dots, f_m^{frg} \rangle \quad (3.16)$$

Estos atributos caracterizan el material del plagio que es encontrado de forma literal sin muchas modificaciones. La existencia de alguna modificación como la eliminación o cambio de alguna palabra provoca una ruptura en la evidencia obtenida (las secuencias comunes). Estas rupturas para la representación, hasta ahora propuesta, son igual de graves si separan al texto común por una palabra o si las alejan de un extremo al otro del documento. Esto es indeseable pues las secuencias que estén cercanas son de mayor evidencia, pues es posible que los cambios sean solo para ocultar el plagio. En este sentido es que propusimos la siguiente, y la última, innovación del método propuesto, con la que posteriormente obtendremos cadenas de texto que contienen mejor evidencia que las secuencias comunes.

3.4 Generalización del concepto del texto reutilizado

Las secuencias comunes, explicadas en el punto 3.1.1, son coincidencias exactas entre los documentos. Estas secuencias son adecuadas para encontrar todos los extractos plagiados sin mayor cambio, pero en muchas ocasiones el plagio está ofuscado¹⁶ para no ser descubierto fácilmente. La introducción u omisión de alguna palabra, e incluso, el cambio de su orden, son operaciones habituales para disfrazar el plagio.

En las secciones anteriores se ha propuesto disminuir el efecto de las secuencias que son falsos positivos al describirlas, usando atributos que aportan información adicional, su fragmentación y distinción. Aquí se propone generalizar el concepto de texto reutilizado con respecto al usado en las secuencias comunes, la intención de esto es bajar la tasa de falsos negativos en las secuencias encontradas como reutilizadas y así poder diferenciar mejor las porciones plagiadas de aquellas que no lo son, y en consecuencia obtener una mayor eficacia en la DAP.

El texto reutilizado es el texto¹⁷ que se ha tomado de otros documentos para conformar el documento sospechoso. En el texto reutilizado es donde radica la mayor parte del material plagiado. Y el concepto del texto reutilizado es donde está la evidencia. Las secuencias comunes es una de las mejores formas en las que se ha podido extraer el texto reutilizado, pero se sabe que no todo lo que se obtiene por las secuencias comunes es texto reutilizado y principalmente no todo el texto reutilizado se puede obtener a partir de las secuencias comunes; esto debido, fundamentalmente, a que el texto reutilizado es modificado para ocultar su origen ilegítimo y para adecuarlo al contexto en el que es implantado.

¹⁶ Utilizamos “ofuscado” como la traducción directa de del termino ingles *obfuscation*, muy utilizado en la literatura de detección automática de plagio. Por “plagio ofuscado” nos referimos al plagio que ha sido modificado para ocultarlo y de esta manera oscurecer o turbar el caso de plagio.

¹⁷ No todo texto reutilizado es un plagio, pues si se da la referencia y el crédito correspondiente, el texto es reutilizado legalmente, pero para fines de la DAP se considera que todo texto reutilizado es parte de un plagio.

La generalización del concepto del texto reutilizado se refiere a la obtención de porciones de texto, que llamaremos cadenas, que son identificadas dentro del concepto de texto reutilizado debido a la flexibilización de la noción de la secuencia.

La generalización del texto reutilizado debe ser adecuada para la tarea y el propósito particular para el que quieren ser usadas, en este caso para la DAP. Aquí, para la obtención de las cadenas, porciones que serán consideradas como texto reutilizado, se ocupa una valoración que llamamos *Índice de Reescritura*. El *Índice de Reescritura* evalúa el grado de reescritura que tiene cada palabra al considerar su vecindad; para ello, utiliza una estrategia de búsqueda de las palabras reutilizadas del documento fuente, estrategia que forma parte de la propuesta de la tesis.

3.4.1 Índice de Reescritura

El *Índice de Reescritura* (*IRe*) tiene por objeto determinar cuánto se asemeja un punto particular del texto sospechoso al documento fuente entero; esta semejanza se puede describir como una “cercanía” entre los textos. Las secciones reescritas serán las áreas del texto con mayor cercanía.

En cada punto i del texto, es decir en cada palabra w_i del documento sospechoso D^s , el *IRe* evalúa qué tan “cerca” está del documento fuente D^f . Para determinar cuán “cerca” están los documentos en el punto i se utiliza una estrategia de búsqueda diseñada *ad hoc* para esta función. El *IRe*, básicamente, califica la distancia como la dificultad de encontrar la palabra w_i en el documento fuente D^f . Evidentemente, para hacer esto, es necesario buscar w_i de una manera inteligente.

La estrategia de búsqueda indaga primero la posición en donde debería de estar w_i para que se tratara de un plagio literal; posteriormente se investiga si se encuentra en posiciones en donde estaría si existiera un cierto grado de reescritura. La búsqueda se dirige desde las posiciones que significarían un mayor grado de reescritura hasta anular la posibilidad de reescritura. Para establecer esta guía, es necesario considerar las posibles operaciones de reescritura y establecer cuál aleja más el texto sospechoso de la fuente.

Las operaciones comunes en la reescritura en los pasajes plagiados [14, 15, 36] son:

- I. Eliminación de palabras
- II. Reordenamiento de la oración (ejemplo de esto es cuando se pasa de voz activa a pasiva);
- III. Inserción de palabras;
- IV. Cambio de palabras (principalmente por sinónimos).

En el *IRe* se establece que la eliminación de un cierto número de palabras es la operación que modifica en menor grado el original; el reordenamiento de palabras es el segundo grado de modificación. La inserción de palabras es el último grado de modificación, considerado como tal y ocupa el último grado de modificación; la inserción de palabras implica la introducción de palabras que no se encuentran en el documento fuente y, por tanto, palabras que parecen ajenas al dominio semántico de la fuente. El cambio de palabras se modela como una eliminación seguida de una inserción y entonces es calificada en el mismo grado de modificación que la inserción simple, siempre y cuando, al igual que la inserción, la palabra insertada sea ajena al léxico del documento sospechoso.

La aplicación de operaciones indiscriminadamente puede llegar al extremo de identificar dos textos totalmente independientes, como el borrado de todas las palabras del primero, seguida de la inserción de todas las del segundo; con el fin de poner restricciones al posible alcance de estas operaciones, en el *IRe*, se define un límite espacial a estas operaciones. Todas las modificaciones que son tomadas como cambios por reescritura deben de actuar dentro de la una región del texto llamada vecindad. La vecindad permite diferenciar las modificaciones realizadas a un pasaje plagiado de la concatenación de distintos pasajes plagiados (de diferentes lugares de la fuente).

El *IRe* considera estas operaciones para la formación de cadenas, flexibilizando las condiciones con las que se formaban las secuencias comunes. En lugar de sólo identificar el texto reutilizado como las secuencias de palabras contiguas copiadas literalmente llamadas secuencias comunes, el *IRe* permite los cambios de orden, y no es necesario que las palabras estén de forma consecutiva. Se sigue manteniendo una observación de la secuencia

de las palabras (ya no como un requisito) y para ello define tres niveles en los que se observa el orden: la secuencia estricta o micro secuencia, la secuencia local y la macro secuencia. La frontera entre la secuencia local y la macro secuencia se determina a través de la dimensión de la vecindad. Mantengamos en mente que la vecindad es tan sólo una región del texto con una dimensión particular. Posteriormente abundaremos en la vecindad, pero antes explicaremos estos tres niveles en los que se presta atención al orden de las palabras.

- La secuencia estricta se refiere a la condición de secuencia que normalmente forma las secuencias comunes, es decir, al orden consecutivo de las palabras.
- La secuencia local es la sucesión que persiste dentro de un margen (definido por la vecindad), sin necesidad de que la conservación del orden sea consecutiva, es decir, entre las palabras que están respetando la secuencia puede haber otras palabras, siempre y cuando el número de ellas no supere un cierto valor.
- La secuencia global es el orden de los elementos a gran distancia, ya no a un contexto particular como la secuencia local sino a la sucesión de palabras considerando todo el documento. Este es el orden de las palabras cuya posición, aunque está en la secuencia adecuada, se encuentra a una gran distancia que supera el margen establecido por la secuencia local.

La estrategia de búsqueda es orientada utilizando la vecindad como una herramienta para que se investiguen primero las áreas del documento fuente en el que se cree que las palabras deberían de estar, si se tratara de un caso de plagio.

La vecindad es una ventana V que almacena v palabras del documento fuente cuya posición, en el documento, está definida por la posición central llamada *foco* que se encuentra exactamente a la mitad de la ventana. El tamaño de la ventana debe de ser siempre impar para que el número de palabras a cada lado del *foco* sea simétrico; así existirán $\frac{v-1}{2}$ palabras a la izquierda y a la derecha del *foco* dentro de V . El objetivo de la vecindad es centrar la búsqueda de las palabras w_i del D^s en una región particular del documento fuente D^f . Esta región es la cubierta por V ; la búsqueda inicia en el *foco* de V que debe estar en la posición adecuada para verificar si w_i es parte de una secuencia

común; si no se encuentra la palabra en el *foco*, inmediatamente después se busca primero en el lado derecho de la ventana V^+ , que son las palabras almacenadas en la parte de la ventana a la derecha del *foco*, si aún no se ha encontrado la palabra, se prosigue con la búsqueda en el otro lado de la ventana, la del lado izquierdo del *foco* V^- . Si se sigue sin encontrar, se realiza la búsqueda afuera de la ventana, en todo el documento, primero a la derecha de la V y luego a la izquierda.

Al realizar esta búsqueda de w_i , una palabra de D^S , en D^F , se determina en qué posición está con respecto al *foco*. La primera comparación que se realiza es de w_i con la palabra contenida en el *foco*, si existiera esta coincidencia, entonces estaríamos haciendo crecer una secuencia común.

Cada palabra recibe el valor del *IRe* que depende de la posición con respecto a V en la que se encontró. La figura 3.1 es una representación gráfica de la vecindad V en la que se muestran los valores del *IRe* que se asignan cuando se halla a w_i en las diferentes posiciones.

	Vecindad V				
Posición	A la izquierda de V	V^-	<i>foco</i>	V^+	A la derecha de V
Valor del índice de reescritura	c_5	c_3	1	c_2	c_4

Figura 3.1 Es la representación de la ventana y de valor que le asigna el *IRe* al encontrar la palabra en alguna de las posiciones indicadas en la figura.

Las condiciones de reescritura que implica encontrar a w_i en alguna de estas áreas son:

1. Coincidencia exacta. w_i es encontrada justo en F ; ésta implica que la palabra es parte de un plagio copiado de forma literal.
2. Coincidencia con eliminación de palabra, se encuentra w_i en V^+ . Aquí, el tamaño de la vecindad determina el número máximo de palabras que se permite borrar, sin perder esta condición de reescritura.

3. Reordenamiento de palabras es cuando se halla a w_i en V . La vecindad determina qué tan lejos pudo ser movida la palabra para juzgar la reescritura como un reordenamiento.
4. Coincidencia ligera de la macro secuencia. Es cuando w_i fue encontrado fuera de V en la parte del documento a la derecha de V o adelante (considerando el orden de lectura) de V ; a partir de ahora, a esta área del documento se le identificará como D_+^F . Esta condición implica que existe una coincidencia de orden débil con el documento fuente, pues las palabras, aunque distantes, están en el mismo orden.
5. Coincidencia ligera. Aquí w_i se halla a la izquierda de V , de igual modo nombramos esta región D_-^F . Esta coincidencia es aún más débil que la anterior, pues sólo implica que la palabra también fue usada en el documento fuente, pero ni siquiera parece ser usada en secciones similares. Rompe con la macro secuencia.

La lista de condiciones está ordenada de la copia exacta a la nula similitud, de la máxima cercanía a la distancia más grande posible; cada condición tiene un valor numérico que indica la cercanía con el texto fuente. El IRe es una medida normalizada por lo que la copia exacta, condición 1, tiene el valor máximo de 1 y la nula similitud, cuando la palabra no se encuentra en el documento fuente, tiene el mínimo 0. El resto de los valores tiene que conservar su orden; el valor para la coincidencia 2 será la constante c_2 , c_3 será para la condición 3, c_4 para la 4 y c_5 para la última condición 5. Estas constantes deben cumplir las desigualdades 3.17.

$$1 > c_2 > c_3 > c_4 > c_5 > 0 \quad (3.17)$$

3.4.2 Algoritmo de búsqueda para la obtención del IRe .

Para obtener el valor de IRe para cada palabra w_i del documento sospechoso D^s es necesario buscar cada w_i en el documento fuente D^f ; para ello se ocupa un algoritmo de búsqueda diseñado específicamente para la obtención del IRe , que explora las áreas del texto con más potencial de contener a w_i , si ésta fuera parte de un texto reutilizado.

El algoritmo utiliza el concepto de vecindad V de la sección anterior, en ella se mantiene la posición en donde se encontró la última coincidencia que forma parte del texto reutilizado y también el contexto de esa coincidencia.

A lo largo de la búsqueda, V se va recorriendo a nuevas posiciones; este mecanismo es llamado alineamiento de V . La forma en la que se recorre no siempre es igual; depende de las condiciones y el lugar en donde se encuentra la próxima coincidencia.

El alineamiento de V se realiza en 6 formas distintas y depende del área en que se encuentra la coincidencia; se ocupan las áreas previamente definidas a partir de V que son *foco*, V^+ , V^- , D_+^F y D_-^F .

- La coincidencia es encontrada en *foco*; aquí la palabra justo donde debería de estar para mantener una secuencia continúa (como en las secuencias comunes). El único alineamiento necesario es recorrer V una palabra a la derecha, de modo que ahora *foco* está en el lugar en donde debería estar la próxima palabra, para continuar creciendo la secuencia.
- La coincidencia sucede en alguna de las posiciones de V^+ , es cuando se han saltado algunas palabras del D^f en el reuso del texto. El alineamiento realizado aquí es recorrer V el número de palabras necesarias para que *foco* quede apuntando a la próxima palabra después de la coincidencia hallada. Este alineamiento tiene el fin de preparar la próxima búsqueda para encontrar una coincidencia en secuencia continua.
- La coincidencia se halla en V^- , esto indica que hay un posible cambio de orden. En este caso V se mantiene exactamente igual, pues no se sabe si en la siguiente palabra que se buscará también estará en un orden distinto o si recuperará la secuencia desarrollada en el documento fuente.
- La coincidencia es encontrada en D_+^F y es pequeña. En esta circunstancia existen dos posibles alineamientos. La decisión de cuál de ellos se utiliza depende del tamaño de coincidencia que se halla en D_+^F ; si la coincidencia, como en este caso, es pequeña, significa que no existe propiamente una secuencia y solamente hay indicios de que las palabras son utilizadas en ambos documentos. En este caso

V se mantiene exactamente igual. La otra circunstancia, cuando la coincidencia en D_+^F es más grande se explicará en el siguiente punto.

- La coincidencia grande sucede en D_+^F , al aparecer una coincidencia de un tamaño considerable en una parte posterior del texto, se intuye que el plagiador ha juntado pasajes distanciados en el documento original. En este caso el alineamiento de V es un poco más complejo que en los casos anteriores; debido a que aquí hay una mezcla de pasajes es importante introducir en V el nuevo pasaje que se está reutilizando, pero también es conveniente mantener aquel que en el que se estaba reutilizado el texto antes de este salto en el documento. Para lograr este compromiso se realizan las siguientes operaciones:

1. Se coloca en V las palabras que se encontraban en V^+ .
2. En *foco* y V^+ se introducen las palabras que continúan la coincidencia mayor encontrada en D_+^F de manera que en el *foco* se encuentra la continuación de la coincidencia.

- La coincidencia se halla en D_-^F , cuando esto sucede significa que la palabra que se buscaba está en el documento fuente, pero está en un área del documento que ya se había descartado como posible fuente del texto reutilizado. En esta situación se deja a V sin cambio alguno.

Si la palabra que se está buscando no está en D^f , entonces no se hace ningún alineamiento especial a V .

El algoritmo que se sigue para evaluar el documento se presenta en la tabla 3.2; tiene una etapa de inicialización. Se posiciona por primera vez a V en el documento fuente D^f en la primera coincidencia entre los documentos. Posteriormente se va tomando cada palabra del documento sospechoso D^s y se va buscando en el D^f empleando el algoritmo de búsqueda de la tabla 3.3 que al determinar su ubicación calcula el *IRe* de las palabras y da el alineamiento apropiado a V .

Algoritmo de evaluación de reescritura en D^s

Algoritmo de reescritura(D^f, D^s)

Inicialización(V, w_i);

Hasta el fin de (D^s)

búsqueda (w_i);

$w_i \leftarrow w_{i+1}$;

Fin

Tabla 3.2 Algoritmo a través del cual se evalúa la reescritura de D^f en D^s .

Algoritmo de Búsqueda

búsqueda (w_i)

Si w_i *está en foco*

$V \leftarrow V[+1]$; // avanza V una posición en D^f

$IRe(w_i) \leftarrow 1$

Salir

Si w_i *está en* V^+

$V \leftarrow V(1 + \text{posicion}_{D^f}(w_i))$; // avanza V a la posición en donde está w_i en D^f

$IRe(w_i) \leftarrow c_2$

Salir

Si w_i *está en* V^-

$V \leftarrow V$; // V se queda exactamente igual

$IRe(w_i) \leftarrow c_3$

Salir

Si w_i *está en* D_+^f

$k = 0$;

Mientras

$((w_{i+k} = D_+^f[\text{posicion}_{D^f}(w_i) + k]) \wedge (k \leq \text{tamaño de coincidencia grande}))$

$k = k + 1$;

Si ($k \geq \text{tamaño de coincidencia grande}$)

Si w_i *está en* D_+^f

$V^- \leftarrow V^+$; // se recorre lo que tiene V^+ a V^-

$\text{foco} \leftarrow D_+^f[1 + \text{posicion}_{D^f}(w_i)]$; // se introduce en **foco** lo que debería de seguir según D^f

$V^+ \leftarrow D_+^f \langle 2 + \text{posicion}_{D^f}(w_i) \rangle;$ $IRe(w_i) \leftarrow c_3$ <p>Salir</p> <p><i>Si w_i está en D_-^f</i></p> $V \leftarrow V; \quad // V \text{ sin cambio alguno}$ $IRe(w_i) \leftarrow c_4$ <p>Salir</p> $IRe(w_i) \leftarrow 0$ <p>Fin</p>	<p>// se introduce en V^+ el resto del pasaje grande que se está encontrando como reutilizado</p>
--	--

Tabla 3.3 Algoritmo usado para buscar la palabras de D^s en D^f y determinar el **IRE** inherente a su búsqueda.

3.4.3 Las cadenas de texto reutilizado

A través del *IRe* se obtienen las porciones de texto que son consideradas como reutilizadas, a las porciones de texto continuas en el D^s las llamamos cadenas c , pues sus enlaces en el D^f no son consecutivos.

Al recorrer el documento sospechoso se obtiene el *IRe* de cada una de sus palabras, se conforman las cadenas con todas las secuencia de palabras que tengan valores de *IRe* superiores a c_3 .

La evaluación de las palabras se hace sobre todo el documento sospechoso D^s y la búsqueda se realiza en el documento fuente D^f . Por tanto el procedimiento de la obtención de las cadenas no es un procedimiento simétrico; esto es comprensible ya que el análisis del plagio debe considerar qué documento pudo haber sufrido las operaciones de reescritura para ser integrado al plagio. De tal modo que se procesa de diferente manera el documento original (susceptible de las modificaciones) que el documento sospechoso (receptor de material plagiado).

El conjunto de todas las cadenas, c , de texto reutilizado obtenidas con el *IRe* es denominado como Ψ^c en forma análoga con el anterior conjunto de las secuencias comunes Ψ . En el método propuesto se utiliza Ψ^c en lugar de Ψ como evidencia a

caracterizar, por lo que en las siguientes sesiones remarcará la forma en la que los atributos son calculados para Ψ^c .

3.4.4 Los atributos de las cadenas de texto reutilizado

Los atributos de fragmentación y distinción son calculados sobre Ψ^c , el conjunto de todas las nuevas secuencias de palabras, llamadas cadenas c , que fueron agrupadas gracias al criterio explicado anteriormente.

En el pesado de la distinción, la expresión 3.18 no tiene sentido para las cadenas, pues aquí la secuencia completa no es una unidad que sea compartida entre los documentos, por lo que el pesado usado es simplemente 3.19.

$$\frac{1}{e^{freq(s_i, D^s) - 1}} \quad (3.18)$$

$$distinción_c(c_i) = \prod_{k=1}^{|c_i|} \frac{2}{freq(w_k^{c_i}, D^s) + freq(w_k^{c_i}, D^f)} \quad (3.19)$$

Con el valor de IRe hemos definido un nuevo atributo que utiliza el IRe a nivel global. Este nuevo atributo f^{IRe} es la suma de los IRe de todas las palabras del documento sospechoso; esta medida permite valorar la reescritura global en todo el documento sospechoso. La expresión matemática con la que se calcula f^{IRe} es la (3.20) en donde $IRe(w_i)$ es una función que permite obtener el valor del IRe de la palabra w_i .

$$f^{IRe} = \sum_{w_i \in D^s} \frac{IRe(w_i)}{|D^s|} \quad (3.20)$$

El atributo f^{IRe} resume la proporción del texto sospechoso que es reconocida como texto reutilizado, por esto en los atributos de fragmentación se prefirió usar la cantidad de secuencias encontradas según el tamaño ahora sin sopesar el tamaño. Por tanto,

las ecuaciones 3.9 y la del atributo que aglutina las secuencias de tamaño mayor a m 3.11, al ser utilizadas ahora para caracterizar las cadenas, adquieren la forma 3.21 Y 3.22

$$f_i^{frag(\Psi^c)} = |\{c_j: c_j \in \Psi^c \wedge |c_j| = i\}| \quad (3.21)$$

$$f_m^{frag(\Psi^c)} = |\{c_j: c_j \in \Psi^c \wedge |c_j| \geq m\}| \quad (3.22)$$

Finalmente los atributos de distinción al caracterizar Ψ^c , ecuación 3.23 y ecuación 3.24, solo se diferencia con respecto a sus análogos de las secuencias comunes (ecuación 3.14 y 3.15) en el uso de la distinción acotada a las cadenas (ecuación 3.19).

$$f_i^{dist(\Psi^c)} = \sum_{\{c: c_j \in \Psi^c \wedge |c_j|=i\}} distinción_c(c_j) \quad (3.23)$$

$$f_m^{dist(\Psi^c)} = \sum_{\{c_j: c_j \in \Psi^c \wedge |c_j| \geq m\}} distinción_c(c_j) \quad (3.24)$$

La representación completa que describe el conjunto de cadenas Ψ^c , que caracteriza a los casos de plagio, se obtienen de mapear Ψ^c a los tres conjuntos de atributos para las cadenas (3.23): los atributos de fragmentación $f_i^{frag(\Psi^c)}$, los atributos de distinción $f_i^{dist(\Psi^c)}$, y el atributo de evaluación global de la reescritura f^{IRe} .

$$\Psi^c \rightarrow \langle f_1^{frag(\Psi^c)}, f_2^{frag(\Psi^c)}, \dots, f_m^{frag(\Psi^c)}, f_1^{dist(\Psi^c)}, f_2^{dist(\Psi^c)}, \dots, f_m^{dist(\Psi^c)}, f^{IRe} \rangle \quad (3.23)$$

4. Evaluación

4.1.1 Disponibilidad de los *corpus* de Plagio.

Lamentablemente no existen muchos *corpus* para evaluar la DAP. Se han generado algunos de manera sintética, donde algoritmos automáticos deciden aleatoriamente la cantidad y longitud del texto que será insertado en un documento anfitrión, que se convertirá en el sospechoso [36,57]. El problema con estos *corpus* es que utilizan un modelo que solo mezcla las partes de los documentos, haciendo una ingenua simplificación del proceso de plagio, principalmente del plagio con alta reescritura, que es justamente el problema más complejo de la DAP. Los casos de plagios generados por estos medios, obviamente, no son muy apegados a los realizados por humanos. En primer lugar, en los plagios humanos existe una mayor cohesión temática entre los documentos sospechosos y las fuentes, pues las fuentes son elegidas conscientemente; otro aspecto es que, obviamente, el tipo de modificaciones y ofuscamiento del plagio es mucho más fino que cuando se hace sintéticamente.

Hay que tener en cuenta que estos *corpus* sintéticos fueron creados primordialmente para la búsqueda de plagio y no para detección de plagio, por lo que las complicaciones en esa tarea está más orientada la complejidad de trabajar con cantidades de datos y no a la de manejar los detalles más finos del lenguaje como se requiere en la DAP.

Dentro de los *corpus* de plagio humano, no todos están disponibles, ya que, como lo indica [11], existen cuestiones éticas y legales¹⁸ que evitan el poder liberar algunos de estos *corpus*. Afortunadamente dentro de dos *corpus* liberados que, con previa autorización del autor, se han podido utilizar para evaluar el método propuesto: el *corpus* METER [58] y el *corpus Plagiarised Short Answers* [59]. El *Corpus* METER es un corpus que no es propiamente para la DAP, pero que en varios trabajos se ha tomado como otro posible corpus de referencia, y *Corpus Plagiarised Short Answers* constituye un verdadero corpus de Plagio. En las siguientes secciones explicaremos más detalles sobre estos dos corpus de evaluación.

4.1.2 *Corpus* METER

El *corpus* METER fue creado para medir la reutilización del texto de los cables en la redacción de los periódicos. Esta reutilización no conforma propiamente un plagio, pues los departamentos de redacción de los periódicos pagan cuotas a las agencias de noticias para poder utilizar libremente todos estos materiales. Si los periódicos no tuvieran permiso de utilizar los contenidos de los cables muchas noticias serían plagios de los cables.

El *corpus* METER está compuesto por 1 717 documentos que se encuentran organizados en 2 grandes grupos: 944 son Noticias publicadas¹⁹ y 773 son cables de noticias de la agencia de noticias PA²⁰. Está organizado de modo que cada noticia tiene su conjunto de cables, los cables y la noticia de estos subconjuntos abordan exactamente el mismo hecho. Cada noticia fue analizada y etiquetada manualmente por un reportero experto que analizó si la noticia contenía material que indicaba la utilización de alguno de

¹⁸ La existencia de material que incrimina a los alumnos, es una de las razones por la que no liberan algunos corpus de plagio.

¹⁹ Para la recolección de noticias fueron monitoreados las secciones de leyes y de negocios de 7 periódicos británicos.

²⁰ *Press Association*, UK.

los cables. Las categorías con las que se etiquetó el corpus son: “*No derivado*”, “*Parcialmente derivado*” y “*Totalmente derivado*”. “*No derivado*” implica que la noticia fue escrita de forma independiente a los cables; “*Parcialmente derivado*” son las noticias que utilizan alguna información de los cables, pero también existe información que es independiente. Por último, las noticias que utilizan sólo la información de los cables sin ningún material original se encuentran en la categoría de “*Totalmente derivado*”.

Para la evaluación de los métodos de la DAP se consideró a las noticias como los documentos sospechosos y a los cables como las fuentes; hemos utilizado únicamente un subconjunto del *corpus* METER. Esta selección está constituida por todas las noticias que cuentan exclusivamente con un cable (es decir, una sola fuente); la selección fue necesaria debido a que las noticias que cuentan con más de un cable no indican de cuál de los cables es que la noticia fue derivada.

De esta manera, el subconjunto donde se probaron los métodos propuestos y los métodos de referencia, está constituido por 253 pares de documentos sospechoso-fuente; se consideraron los documentos con las etiquetas “*Parcialmente derivado*” y “*Totalmente derivado*” como los casos de plagio y a los “*No derivado*” como no plagiados.

4.1.3 Corpus Plagiarised Short Answers

Constituye el único *corpus* de verdadero plagio humano disponible en la actualidad. Para la construcción de este corpus se les pidió a 19 estudiantes contestar 5 preguntas relacionadas con su área de estudio (ciencias computacionales). Se pidieron respuestas de entre 200 y 300 palabras de longitud; cada una de las preguntas incluía instrucciones específicas para que fueran contestadas según una estrategia particular de plagio o sin cometerlo. En los casos de plagio se les proporcionó la fuente de la cual se extraería la información y en los casos de no plagio se les suplicó no utilizar esa fuente (ni como referencia).

El *corpus* está constituido por 100 documentos: 5 fuentes, una para cada una de las 5 preguntas, y 95 Textos, que respondían a una de las 5 preguntas. Las categorías de los documentos de respuestas son:

- No plagiados. Se les pidió a los participantes que escribieran estos documentos sin plagio y sin haber revisado la fuente sugerida para los documentos plagiados.
- Copias cercanas. Los participantes debían responder a la pregunta a través de simples copias del texto de referencia; debían elegir el material relevante para contestar la pregunta y las modificaciones para que constituyeran un texto legible.
- Revisión ligera. Aquí se les pidió a los participantes contestar las preguntas basándose en el texto de la fuente proporcionada, haciendo algunas sustituciones de palabras y reestructurando algunas oraciones; pero se les indicó enfáticamente que no se alterara severamente la estructura de la información de la fuente.
- Revisión severa. Las respuestas en esta categoría deben de tener el mismo sentido que el texto de la fuente, pero aquí se pidió explícitamente que se realizara una nueva redacción de la información, que propusieran un nuevo orden de las ideas, que usaran otras palabras y estructuras para reexpresar la misma información.

Para la evaluación de los métodos de la DAP se integraron los pares de documentos; cada respuesta proporcionada por los participantes se tomó como documento sospechoso y la respuesta de referencia que se les entregó como fuente. En el caso de la evaluación binaria de la DAP se tomaron todas las estrategias de plagio, copia cercana y las revisiones ligera y severa, como documentos plagiados.

4.2 Métricas de evaluación

Para evaluar la eficacia de la DAP se ha empleado el conjunto de métricas disponibles de la clasificación automática. La DAP puede ser vista como una clasificación en donde los objetos a clasificar son los documentos sospechoso y las categorías en las que se van a clasificar determinan la existencia del plagio y el tipo de plagio. Dependiendo cuál sea el interés particular que se tenga en la clasificación, ésta puede ser binaria con las categorías

de *plagio* y *no plagio* o multiclase, donde cada categoría especifica un grado de reescritura o un tipo de plagio. Es conveniente, además, saber cuando existe plagio, qué tipo de plagio es, pues esto nos permitirá realizar una verificación más rápida o descartar la revisión de algunas de las categorías. En el caso del *corpus Plagiarised Short Answers* fue posible evaluar la predicción del tipo de plagio.

Para la evolución de un clasificador se tiene que cotejar la predicción del clasificador y la clase real de los objetos de evaluación. Para este análisis se utiliza una matriz de confusión; la tabla 4.1 presenta una matriz de confusión. En las columnas se tiene la clase que el clasificador ha predicho y, en las filas, las clases a las que realmente pertenecen los objetos. En esta matriz de confusión se registra el número de aciertos y errores (por cada clase) que tuvo el clasificador al ser evaluado con un conjunto de objetos particular. TP (*true positives*) son el número de aciertos de la clase positiva y TN (*true negatives*) los de la clase negativa. Por otro lado FP (*false positive*) es el número de errores que tiene la clase positiva y FN (*false negative*) son de la clase negativa. Los errores por clase se refieren a los objetos que fueron predichos como cierta categoría por el clasificador, cuando en realidad son de otra.

Teniendo la matriz de confusión (tabla 4.1):

		Predicción	
		Positiva	Negativa
Real	Positiva	<i>TP</i>	<i>FN</i>
	Negativa	<i>FP</i>	<i>TN</i>

Tabla 4.1 Matriz de confusión para clasificación binaria

Con estos datos, normalmente, se calculan dos medidas: precisión (ecuación 4.1) y recuerdo (ecuación 3.1). La precisión, P , es una medida que indica cuántos de los objetos fueron correctamente predichos en cierta clase. El recuerdo indica cuántos de los objetos de una cierta clase fueron reconocidos como tales. En el caso de tener dos clases tenemos que:

$$P_{Positiva} = \frac{TP}{(TP+FP)} \quad (4.1)$$

$$R_{Positiva} = \frac{TP}{(TP+FN)} \quad (4.2)$$

$$P_{Negativa} = \frac{TN}{(TN+FN)} \quad (4.3)$$

$$R_{Negativa} = \frac{TN}{(TN+FP)} \quad (4.4)$$

Debido a que cada una de las medidas atiende a características distintas, ambas deseables, y para manejar el compromiso entre ellas, se ha utilizado la F_1 -*measure*, que involucra ambos aspectos. Al igual que para el recuerdo y la precisión existe una F_1 -*measure* para cada clase (ecuación 4.5 y 4.6), y para tener un desempeño global particular en la DAP, se ha utilizado el promedio de F_1 -*measure* de las clases (ecuación 4.7) [12].

$$\begin{aligned} f_1 - measure_{Positiva} &= \frac{2 P_{Positiva} \cdot R_{Positiva}}{P_{Positiva} + R_{Positiva}} \\ &= \frac{2 TP}{(TP + FP) + (TP + FN)} \end{aligned} \quad (4.5)$$

$$\begin{aligned} f_1 - measure_{Negativa} &= \frac{2 P_{Negativa} \cdot R_{Negativa}}{P_{Negativa} + R_{Negativa}} \\ &= \frac{2 TN}{(TN + FN) + (TN + FP)} \end{aligned} \quad (4.6)$$

$$f - measure_{Promedio} = \frac{f_1 - measure_{Positiva} + f_1 - measure_{Negativa}}{2} \quad (4.7)$$

Otra medida de desempeño global es la *exactitud* (ecuación 4.8), que es el número de aciertos global del clasificador.

$$exactitud = \frac{TP + TN}{TP + FP + TN + FN} \quad (4.8)$$

4.3 Condiciones de los experimentos

4.3.1 El clasificador

En todos los experimentos se ha utilizado el clasificador Naive Bayes; esto, debido las mismas consideraciones hechas en [12]: es uno de algoritmos más sencillos para la clasificación y ha mostrado buenos resultados en otras tareas del procesamiento de textos.

El clasificador Naive Bayes [56] pertenece al grupo de los clasificadores probabilísticos, calcula la probabilidad de que un documento d_x pertenezca a la clase C_i . La probabilidad de que el documento d_x pertenezca a la clase C_i (4.9) se puede calcular de la siguiente forma:

$$P(C_i|d_x) = \frac{P(C_i)P(d_x|C_i)}{P(d_x)} \quad (4.9)$$

Donde $P(C_i)$ es la probabilidad de la clase C_i que se puede calcular como el número de documentos de C_i entre el total de documentos; $P(d_x)$ es la probabilidad del documento d_x pero es posible eliminar este término gracias a que lo único que nos interesa es saber a qué clase pertenece d_x , es decir, para cuál C_i , $P(C_i|d_x)$ tiene el valor máximo. Por último $P(d_x|C_i)$ es la probabilidad del documento dado que pertenece a una la clase C_i y para su cálculo se usa el conjunto de atributos $\langle f_1, f_2, \dots, f_{m-1}, f_m \rangle$ que describen al documento d_x . Este clasificador supone que todos los atributos son independientes respecto a la clase, hecho que, en general, no es cierto, por lo que recibe el nombre de *naive* o ingenuo; pero esta simplificación permite calcular $P(d_x|C_i)$ mediante la expresión 4.10.

$$P(d_x|C_i) = \prod_{k=1}^{k=m} P(f_k|C_i) \quad (4.10)$$

Finalmente, se tiene que la clase C'_x predicha para d_x es aquella que maximiza la expresión $P(C_i|d_x)$ 4.11.

$$P(C_i|d_x) = P(C_i) \prod_{k=1}^{k=m} P(f_k|C_i) \quad (4.11)$$

Por tanto la clase predicha C'_x para d_x (4.12) es:

$$C'_x = \max_{C_i}(P(C_i|d_x)) \quad (4.12)$$

4.3.2 El método de evaluación

Para evaluar todos los métodos de la DAP se utilizó validación cruzada en k pliegues (*cross k-fold validation*) [56], usualmente utilizada en la evaluación de los métodos de clasificación. Se emplearon 10 pliegues (k=10).

Esta técnica consiste en una segmentación del *corpus* de evaluación en k partes no traslapadas y en realizar por separado las k clasificaciones. Este método permite que todos los ejemplos con los que se cuenta para la evaluación estén al menos una vez en el conjunto de entrenamiento y en el de la evaluación. El método permite aumentar la confianza de que los resultados no dependen del conjunto de evaluación particular ni del conjunto particular de entrenamiento.

4.3.3 Preprocesamiento del *corpus*

En muchas tareas del procesamiento automático de textos, como la clasificación, el agrupamiento, la recuperación de información, etc., se aplican algunos procesos previos a los realizados propiamente para la tarea particular. Estos procesos tienen como principal objetivo el de limpiar y eliminar información que no es útil para la tarea particular; los más habituales son el *stemming* y la eliminación de *stop words*. El *stemming* o lematización es la transformación de las palabras en una expresión cercana a su raíz, llamada lema; la eliminación de *stop words* o palabras vacías es un proceso en el cual se descartan del texto palabras carentes de sentido propio y muy comunes en el lenguaje, como son las preposiciones, conjunciones y algunos otros elementos. En la DAP existen algunos trabajos que dan por hecho que, debido a que es un procesamiento automático, estos procesos deben de hacerse sin cuestionamiento; algunos otros opinan lo contrario es necesario mantener la

forma original del texto, pues la información de las palabras vacías y las formas de las palabras se deben de conservar, ya que ayudarán a reconocer el plagio. Este último enfoque es respaldado por los trabajos que realizan los experimentos en ambos sentidos y encuentran que manejar los textos sin preprocesamiento severo es más eficaz en la DAP [42]. Por tanto, nosotros hemos decidido manejar los textos sin preprocesamiento, excepto el cambio de los signos de puntuación [2] por una etiqueta genérica de puntuación. Esto debido a que algunos efectos de la reescritura son simples cambios en estos signos.

4.4 Métodos de referencia

Se realizaron múltiples experimentos para poder comparar el método propuesto. La selección de estos métodos obedece a dos aspectos fundamentales: el primero es a la revisión bibliográfica, que permitió obtener aquellos que tenían los mejores desempeños; estos métodos han sido llamados “métodos modernos de detección” [5]; el segundo aspecto responde a obtener métodos de referencias que estén en condiciones semejantes con el método propuesto, como es que los métodos de comparación no tengan dependencia de herramientas del lenguaje, como ontologías, diccionarios, etc. A diferencia de los métodos de referencia obtenidos del estado del arte, el método propuesto utiliza más de un atributo. Atendiendo a este último punto, se propuso una nueva versión de uno de los métodos de referencia, en donde se le permite tener más de un atributo y con esto eliminar algún posible sesgo por la cantidad de atributos que con la que pueden describir los casos de plagio en los métodos con los que se realiza la comparación de los métodos de referencia.

Los métodos tomados como referencia son:

1. El cubrimiento con n-gramas, es decir, aquel en el que se fragmentan los documentos sospechoso y fuente en pequeños trozos de n palabras llamados n-gramas para medir la cantidad de traslape (normalizada con la medida *containment*).

2. El cubrimiento con secuencias comunes (o también llamadas *tiles*), en donde de nuevo se mide el traslape que se tiene entre los textos; es decir, se cuentan las palabras de todas las secuencias comunes recolectadas y se expresan en la razón de palabras reutilizadas del total de palabras del documento sospechoso. El método de cubrimiento con secuencias comunes, para eliminar el ruido que puede contener las secuencias pequeñas (como preposiciones comunes o términos temáticos), filtra las secuencias comunes para solo considerar aquellas que son de tamaño mayor o igual a cierto umbral, llamado MML (*Minimum Match Length*).

Para tener un panorama adecuado de la eficiencia de los dos métodos anteriores, el cubrimiento con n-gramas y el cubrimiento con secuencias comunes, se realizaron experimentos con amplio rango de n (1 a 10 para un *corpus* METER y de 1 a 15 para el *Corpus Plagiarised Short Answers*), y con un rango de MML de 1 a 10 (o hasta 15). Cabe aclarar que realizamos estos experimentos en un rango de los parámetros de control tan amplio o incluso más amplio de lo que usualmente se realizan ([12,32]), esto con el objeto de garantizar que se encuentre el punto óptimo de operación, así, como, esclarecer el comportamiento que tiene la eficacia de los métodos al variar estos parámetros.

3. Empleamos una nueva²¹ referencia, el método de conjuntos de atributos de cubrimiento de n-gramas, para obtener una comparación, que al igual que el método propuesto contenga más de un solo atributo y que considere secuencias de diferentes tamaños. Este nuevo método se presenta para realizar una comparación en mejores condiciones de igualdad. En este método se utiliza un conjunto de atributos; cada uno de ellos es el cubrimiento con n-gramas para un cierto valor n, y el conjunto se forma utilizando varios valores de n. En cada experimento se incrementa un nuevo atributo al conjunto de atributos con el que se realiza la DAP, la elección de los atributos es de forma ascendente.

²¹ Este método es propuesto en este trabajo por primera vez, aunque no es el principal método que se presenta en este trabajo, simplemente es un nuevo método que se puede ver como una extensión del cubrimiento con n-gramas. Este método se define aquí para ser un *baseline* más próximo a las características del método propuesto.

4.5 Resultados de los métodos de referencia

Los resultados de los experimentos con cubrimiento de n-gramas con la variación de la n sobre el *corpus* METER se encuentran en la tabla 4.2 y los obtenidos con el *Corpus Plagiarised Short Answers* están en la tabla 4.3. En el caso de los que se realizaron en el *Corpus Plagiarised Short Answers* tienen un rango n de 1 a 15 debido a que en experimentos posteriores fue necesario utilizar un rango mayor al utilizado en el *corpus* METER (1 a 10) para observar el comportamiento a lo largo de la variación del método.

n-grama	<i>f-measure</i>		Promedio de <i>f-measure</i>	<i>Exactitud</i>
	<i>“si plagio”</i>	<i>“no plagio”</i>		
2	0.785	0.563	0.674	0.711
3	0.736	0.553	0.644	0.667
4	0.717	0.574	0.645	0.660
5	0.692	0.569	0.630	0.640
6	0.676	0.565	0.620	0.628
7	0.65	0.545	0.597	0.604
8	0.621	0.531	0.576	0.581
9	0.596	0.530	0.563	0.565
10	0.564	0.517	0.540	0.541

Tabla 4.2 Resultados de la eficacia del método de cubrimientos con n-gramas al variar el parámetro n sobre el *corpus* METER.

n-grama	<i>f-measure</i>				Promedio de <i>f-measure</i>	<i>Exactitud</i>
	“no plagio”	“revisión severa”	“copias cercanas”	“revisión ligera”		
1	0.892	0.359	0.385	0.429	0.516	0.610
2	0.925	0.439	0.571	0.353	0.572	0.652
3	0.911	0.500	0.571	0.375	0.589	0.663
4	0.900	0.465	0.556	0.387	0.577	0.652
5	0.900	0.489	0.632	0.370	0.597	0.673
6	0.902	0.500	0.632	0.308	0.585	0.673
7	0.867	0.465	0.649	0.296	0.569	0.652
8	0.894	0.512	0.611	0.231	0.562	0.663
9	0.864	0.419	0.611	0.174	0.517	0.631
10	0.854	0.372	0.649	0.095	0.492	0.621
11	0.835	0.410	0.649	0.087	0.495	0.621
12	0.809	0.368	0.667	0.091	0.483	0.610
13	0.800	0.316	0.629	0.091	0.459	0.589
14	0.784	0.286	0.629	0.087	0.446	0.578
15	0.776	0.278	0.629	0.00	0.420	0.568

Tabla 4.3 Resultados de la eficacia del método de cubrimientos con n-gramas al variar el parámetro n sobre el *corpus Plagiarised Short Answers*.

Los resultados de la propuesta de los grupos de atributos de los n-gramas, conforme se van agregando los atributos, se encuentran en la tabla 4.4 y la tabla 4.5. En el caso de los experimentos, en el *corpus Plagiarised Short Answers* se tuvieron que hacer en un rango más amplio pues el punto óptimo del parámetro (n=10) estaba en el límite del rango normalmente explorado.

Conjunto de atributos de cubrimiento de n-gramas					
n-gramas incluidos	número de atributos	f-measure		Promedio de f-measure	Exactitud
		“si plagio”	“no plagio”		
{1}	1	0.817	0.493	0.655	0.731
{1-2}	2	0.808	0.628	0.718	0.747
{1-3}	3	0.766	0.611	0.688	0.707
{1-4}	4	0.728	0.598	0.663	0.675
{1-5}	5	0.711	0.587	0.649	0.660
{1-6}	6	0.707	0.584	0.645	0.656
{1-7}	7	0.681	0.578	0.629	0.636
{1-8}	8	0.667	0.570	0.618	0.624
{1-9}	9	0.655	0.569	0.612	0.616
{1-10}	10	0.645	0.564	0.604	0.608

Tabla 4.4 Resultados de la eficacia del nuevo método de conjuntos de atributos de cubrimiento de n-gramas al incrementar el conjunto de los atributos sobre el corpus METER.

Conjunto de atributos de cubrimiento de n-gramas							
n-gramas incluidos	número de atributos	f-measure				Promedio de f-measure	Exactitud
		“no plagio”	“revisión severa”	“copias cercanas”	“revisión ligera”		
{1}	1	0.892	0.359	0.385	0.429	0.516	0.591
{1-2}	2	0.925	0.429	0.606	0.457	0.604	0.673
{1-3}	3	0.95	0.524	0.588	0.412	0.618	0.694
{1-4}	4	0.937	0.512	0.571	0.364	0.596	0.673
{1-5}	5	0.925	0.476	0.571	0.364	0.584	0.663
{1-6}	6	0.914	0.476	0.556	0.387	0.583	0.663
{1-7}	7	0.9	0.465	0.556	0.387	0.577	0.652
{1-8}	8	0.914	0.476	0.611	0.452	0.613	0.684
{1-9}	9	0.914	0.476	0.611	0.452	0.613	0.684
{1-10}	10	0.914	0.545	0.611	0.483	0.638	0.705
{1-11}	11	0.914	0.545	0.611	0.483	0.638	0.705
{1-12}	12	0.914	0.522	0.611	0.37	0.604	0.684
{1-13}	13	0.914	0.522	0.571	0.357	0.591	0.673
{1-14}	14	0.902	0.478	0.571	0.296	0.561	0.652
{1-15}	15	0.902	0.478	0.571	0.296	0.561	0.652

Tabla 4.5 Resultados de la eficacia del nuevo método de conjuntos de atributos de cubrimiento de n-gramas al incrementar el conjunto de los atributos sobre el corpus *Plagiarised Short Answers*.

Los resultados de los experimentos con el cubrimiento con las secuencias comunes o *tiles* se encuentran en la tabla 4.6 y la tabla 4.7. Al igual que en los experimentos de cubrimientos con n-gramas, fue necesario extender el rango de los experimentos en los que se realizaron en el *Corpus Plagiarised Short Answers*.

MML	<i>f-measure</i>		Promedio de <i>f-measure</i>	<i>Exactitud</i>
	“ <i>si plagio</i> ”	“ <i>no plagio</i> ”		
1	0.794	0.391	0.592	0.691
2	0.805	0.549	0.677	0.727
3	0.805	0.548	0.676	0.727
4	0.759	0.571	0.665	0.691
5	0.727	0.576	0.651	0.667
6	0.724	0.584	0.654	0.667
7	0.709	0.580	0.644	0.656
8	0.685	0.570	0.627	0.636
9	0.671	0.562	0.616	0.624
10	0.539	0.648	0.593	0.600

Tabla 4.6 Resultados de la eficacia del método del cubrimiento con secuencias comunes al variar el MML sobre el *corpus* METER.

MML	<i>f-measure</i>				Promedio de <i>f-measure</i>	<i>Exactitud</i>
	“ <i>no plagio</i> ”	“ <i>revisión severa</i> ”	“ <i>copias cercanas</i> ”	“ <i>revisión ligera</i> ”		
1	0.894	0.400	0.385	0.409	0.522	0.621
2	0.914	0.400	0.385	0.465	0.541	0.631
3	0.914	0.400	0.571	0.412	0.574	0.652
4	0.914	0.359	0.556	0.353	0.545	0.631
5	0.900	0.429	0.571	0.364	0.566	0.642
6	0.914	0.512	0.571	0.387	0.596	0.673
7	0.914	0.558	0.595	0.345	0.603	0.684
8	0.914	0.558	0.615	0.370	0.614	0.694
9	0.894	0.488	0.632	0.385	0.599	0.684
10	0.874	0.400	0.632	0.320	0.556	0.652
11	0.844	0.368	0.632	0.250	0.523	0.631
12	0.826	0.333	0.595	0.160	0.478	0.600
13	0.809	0.389	0.595	0.174	0.491	0.610
14	0.809	0.324	0.667	0.000	0.450	0.600
15	0.784	0.250	0.684	0.174	0.473	0.600

Tabla 4.7 Resultados de la eficacia del método del cubrimiento con secuencias comunes al variar el MML sobre el *corpus* *Plagiarised Short Answers*.

Las tablas 4.8 y 4.9 resumen los mejores resultados para ambos *corpus* de todos los métodos de referencia: cubrimiento con n-gramas, los conjuntos de atributos de n-gramas y cubrimiento con secuencias comunes.

Resumen de los mejores resultados de los métodos de referencia					
Método	número de atributos	<i>f-measure</i>		Promedio de <i>f-measure</i>	<i>Exactitud</i>
		“ <i>si plagio</i> ”	“ <i>no plagio</i> ”		
1-grams	1	0.817	0.493	0.655	0.731
2-gramas	1	0.785	0.563	0.674	0.711
{1-2}-gramas	2	0.808	0.628	0.718	0.747
Secuencias comunes con MML=2	1	0.805	0.549	0.677	0.727

Tabla 4.8 Mejores resultados sobre el *corpus* METER.

Resumen de los mejores resultados de los métodos de referencia							
Método	número de atributos	<i>f-measure</i>				Promedio de <i>f-measure</i>	<i>Exactitud</i>
		“ <i>no plagio</i> ”	“ <i>revisión severa</i> ”	“ <i>copias cercanas</i> ”	“ <i>revisión ligera</i> ”		
5-gramas	1	0.900	0.489	0.632	0.370	0.597	0.673
{1-10}-gramas	10	0.914	0.545	0.611	0.483	0.638	0.705
Secuencias comunes con MML=8	1	0.914	0.558	0.615	0.370	0.614	0.694

Tabla 4.9 Mejores resultados sobre el *corpus* *Plagiarised Short Answers*.

4.6 Discusión de los resultados de los métodos de referencia

En las tablas 4.8 y 4.9, que resumen los mejores resultados, observamos que las diferencias entre los métodos de referencias, el cubrimiento, con n-gramas y con secuencias comunes, es de 0.731 a 0.747 en *exactitud* para el *corpus* METER y de 0.67 a 0.69 en el *corpus*

Plagiarised Short Answers. La diferencia entre estos métodos no es muy grande en el *corpus* METER y un poco mayor en *corpus Plagiarised Short Answers*. Pero es digno de mencionar que no es claro cuál de estos dos métodos es la mejor opción, pues resulta mejor el cubrimiento con n-gramas para el *corpus* METER y para el *corpus Plagiarised Short Answers* se desempeña mejor el cubrimiento con secuencias comunes.

En ambos *corpus* el método con la mejor evaluación fue el conjunto de atributos cubrimiento con n-gramas. Este método no es una estrategia antes propuesta en el estado del arte; este método, que resultó el mejor evaluado, constituye el *baseline* propuesto en este trabajo para tener una comparación en mejores condiciones de igualdad. Se ha incluido este método como parte de las referencias para la comparación pues es la forma más sencilla y directa de extender algún método existente para que pueda generar una variedad de atributos que consideren diferentes tamaños de las coincidencias exactas. El éxito de este método, da pauta para pensar que es factible la hipótesis del método propuesto: utilizar atributos exclusivos para cada longitud de las secuencias comunes mejora la eficacia de la DAP.

La mayoría de los mejores resultados en el Promedio de *f-measure* y la *exactitud* coinciden con excepción del cubrimiento con n-grama en el *corpus* METER donde la existe una diferencia que es pequeña comparándola con la diferencia con el resto de los resultados de ese método, pues estos dos resultados son mejores en ambas medidas. La coincidencia en la mayoría de los resultados muestra que ambas medidas son aparentemente buenas para evaluar la DAP.

El cubrimiento con n-gramas en el *corpus* METER resulta tener mejor evaluación en los casos de plagio, sin embargo, con n-grama mayores (al de la mejor exactitud), la eficacia de los casos de no plagio incrementa, teniendo un máximo en n igual a 4. Esto último, indica que los 4-gramas deben ser lo suficientemente grandes para que con su baja existencia se pueda descubrir la inexistencia de plagio; pero no demasiado grandes como para que puedan existir coincidencias que se identifiquen cuando exista plagio. La eficacia de la clase donde si hay plagio, decrementa conforme la n crece, esto es sensato pues los n-gramas grandes serán poco comunes, aún en caso de plagio.

Los resultados de la aplicación de cubrimiento con n-gramas en el *corpus Plagiarised Short Answers* al contrario que en el *corpus* METER, los caso no plagiados, son los que puede clasificar mejor. La siguiente clase mejor clasificada es “*copia cercana*” (lo que es de esperarse) pues en ésta es la que contiene “*copy-paste*” que debe ser fácil de detectar con los n-gramas; en especial, como muestran los resultados, los n-gramas de mayor tamaño clasificaran mejor los casos de “*copia cercana*”. Lo anterior implica que este método clasifica relativamente bien los casos extremos, “*no plagio*” y “*copia cercana*” y es malo para los casos donde entra en juego la reescritura, las clases “*revisión ligera*” y “*revisión severa*”. Dentro de estas dos clases con reescritura al contrario de lo que se pensaría “*revisión severa*” tiene mejor evaluación que “*revisión ligera*”, y “*revisión severa*” que tiene un máximo en su evaluación en una n intermedia (n=8).

El cubrimiento con secuencias comunes en el *corpus* METER tiene una mejor eficiencia en la clase de plagio. Como es de esperarse al ir aumentado el tamaño mínimo de las secuencias comunes consideradas (mediante el filtro de la MML) los casos de plagios son peor clasificados y los de no plagio mejoran su clasificación.

En cubrimiento con secuencias comunes en el *corpus Plagiarised Short Answers* el comportamiento de la clasificación a través de las clases es similar al logrado con cubrimiento con n-gramas, la principal diferencia es en la clase “*revisión severa*” pues su eficacia es superior, seguramente debido a que el filtro debe de cumplir bien su función de eliminación de ruido por términos temáticos.

El método del conjunto de atributos de cubrimiento con n-gramas en el *corpus* METER tiene una mejor eficiencia en la clase “*si plagio*”. Al usar el conjunto, a diferencia de cuando se usan solo los n-gramas separados, la caída de la eficacia de la clasificación de “*si plagio*” es más lenta y la clase “*no plagio*” presenta un máximo más alto en la eficacia de clasificación.

Cuando se utiliza el método del conjunto de atributos de cubrimiento con n-gramas en el *corpus Plagiarised Short Answers*, las clases cuyas eficiencias de clasificación se benefician al tener el conjunto, en lugar de un simple atributo de algún tamaño específico, son las que presentan reescritura, “*revisión ligera*” y “*revisión severa*”.

A lo largo de los experimentos de los métodos de referencia, se observa que los casos de reescritura “*revisión ligera*” y “*revisión severa*” son los casos más difíciles de clasificar.

Encontrar que el mejor método de referencia en la clasificación de los casos difíciles es el de los conjuntos de atributos de n-gramas, apoya la idea de que tener más atributos dedicados a cada uno de los diversos tamaños de las coincidencias exactas, ayuda a caracterizar el plagio y por tanto clasificarlo mejor.

Finalmente, como se ha expuesto en la revisión del estado del arte, el tamaño del n-grama para un buen desempeño no es el mismo en *corpus* de naturalezas distintas, y en el caso de nuestros corpus de evaluación obtuvimos que para el *corpus* METER, los mejores resultados fueron con una n igual a 1 o 2; y para el *corpus Plagiarised Short Answers*, el mejor fue con 5. De igual modo, el filtro del método de cubrimiento de secuencias comunes tuvo diferentes valores óptimos del MML; en el *corpus* METER fue de 3 y en el *corpus Plagiarised Short Answers* de 8. Esta discrepancia en los valores óptimos ratifica la sensibilidad de los métodos a la naturaleza del *corpus*.

4.7 Experimentos de los métodos propuestos

Se realizaron experimentos sobre los dos *corpus* disponibles para evaluar al método propuesto. Los experimentos evalúan las diferentes ideas básicas del método que están en cada uno de los 3 ejes en los que se enfocan las innovaciones del método propuesto (véase capítulo 3, el método propuesto). Estas innovaciones están contenidas en la utilización de cadenas y en los dos grupos de atributos, los de fragmentación y los de distinción. Realizamos experimentos para evaluar la introducción de cada una de estas ideas en el método propuesto.

Los experimentos realizados fueron:

- La utilización únicamente de los atributos de fragmentación (f^{frag}).
- La incorporación del pesado de distinción; por tanto, se utilizaron los atributos de fragmentación y de distinción (f^{frag} y f^{dist}).
- La utilización de las cadenas, usando el índice de reescritura propuesto y su representación, utilizando los atributos de fragmentación y de distinción. ($f^{frag(\psi^c)}$, $f^{dist(\psi^c)}$, y f^{IRe}).

El método propuesto tiene algunos parámetros que es necesario definir; el primero de ellos es el subíndice máximo m de los atributos. Con éste se define el número de atributos que tendrá la representación y, por lo tanto, el número de atributos especializados en longitudes de secuencias particulares. Para definir el valor de m se generan los atributos de fragmentación sobre una partición del corpus con una m grande. Se obtiene la Ganancia de Información (G.I.) de todos los atributos para definir un punto de corte, que será el punto en donde los atributos obtienen ganancias de información relativamente pequeñas. Con este punto de corte se define el nuevo m definitivo.

Se realizó el experimento con el cual se define la m , en cada una de los 10 pliegues de la validación cruzada del método en ambos corpus, en la tabla 4.10 se muestra el promedio y desviación estándar de las G.I. de los atributos en los diez pliegues para ambos corpus (para las tablas de cada corpus por separado consultar el apéndice I). Se tomó para el experimento una $m=50$; sólo se muestran las G.I. de los primeros 6 atributos, pues con ellos es posible observar el comportamiento general y así definir la m apropiada ($m=3$).

Longitud de las secuencias	Promedio de G.I.	Desviación estándar
1	0.467	0.049
2	0.266	0.025
$m=3$	0.206	0.044
4	0.034	0.027
5	0.012	0.037
6	0.172	0.023

Tabla 4.10 Muestra el promedio de la G.I. y su desviación estándar en las 10 particiones hechas de los corpus.

El segundo parámetro a definir es el tamaño de la vecindad considerada en la extracción del *IRe*; para ello se siguió una estrategia similar a la empleada con la definición del valor de la *m*. Se hicieron experimentos con valores de 3, 5, 11, 15, 21 y 25 palabras de longitud de la vecindad y se calculó la G.I. de los atributos f^{IRe} generados para cada valor (se puede ver la tabla de estos valores en el apéndice I); el valor elegido es el del mejor el mejor valor de G.I. el cual fue 5. El resto de los valores definidos son también utilizados en el *IRe* y son las constantes c_2 , c_3 , c_4 y c_5 , para ellas se experimenta con cuatro funciones discretas que cumplieran la condición de desigualdad requerida (4.13).

$$1 > c_2 > c_3 > c_4 > c_5 > 0 \quad (4.13)$$

Las cuatro funciones fueron (4.14), (4.15), (4.16), y (4.17):

$$c_n = \frac{1}{n} \quad (4.14)$$

$$c_n = \frac{1}{n^2} \quad (4.15)$$

$$c_n = \frac{1}{n^3} \quad (4.16)$$

$$c_n = \frac{1}{e^n} \quad (4.17)$$

La función (4.14) genera el atributo f^{IRe} con el un mejor valor de G.I. (en el apéndice I se muestra la tabla de los valores de las G.I.) y como se muestra en (4.18) cumple perfectamente la ecuación (4.13).

$$1 > \frac{1}{2} > \frac{1}{3} > \frac{1}{4} > \frac{1}{5} > 0 \quad (4.18)$$

Estos parámetros se definen experimentado en ambos *corpus* de evaluación (METER y *corpus Plagiarised Short Answers*) y se utilizaron las mismos 10 pliegues empleadas en la evaluación de los diversos métodos (de referencia y los propuestos) para medir la G.I. Finalmente se obtiene el promedio y la desviación estándar del G.I. a través de los pliegues; este procedimiento de elección de los parámetros se eligió para poder

proporcionar un acercamiento a la elección de valores apropiados en un esquema en donde solo se cuente con un pequeño conjunto de entrenamiento²².

4.8 Resultados de los métodos propuestos

En la tabla 4.11 y 4.12 se encuentran los resultados sobre ambos *corpus*; de los 3 experimentos de las principales innovaciones que tiene el método propuesto, el último experimento constituye la evaluación global del método propuesto, en la que se incluyen todas las innovaciones hechas en el método propuesto.

Experimentos de las propuestas y el método completo (3 ^{er})					
Atributos	número de atributos	<i>f-measure</i>		Promedio de <i>f-measure</i>	<i>Exactitud</i>
		“ <i>si plagio</i> ”	“ <i>no plagio</i> ”		
f^{frg}	3	0.851	0.500	0.675	0.770
f^{frg} y f^{dist}	6	0.856	0.553	0.704	0.782
$f^{frg(\psi^c)}$, $f^{dist(\psi^c)}$ y f^{IRe}	7	0.862	0.526	0.694	0.786

Tabla 4.11 Resultados del método propuesto y de sus principales ideas en el *corpus* METER.

Experimentos de las propuestas y el método completo (3 ^{er})							
Atributos	número de atributos	<i>f-measure</i>				Promedio de <i>f-measure</i>	<i>Exactitud</i>
		“ <i>no plagio</i> ”	“ <i>revisión severa</i> ”	“ <i>copias cercanas</i> ”	“ <i>revisión ligera</i> ”		
f^{frg}	3	0.911	0.462	0.424	0.410	0.551	0.631
f^{frg} y f^{dist}	6	0.883	0.486	0.553	0.414	0.584	0.652
$f^{frg(\psi^c)}$, $f^{dist(\psi^c)}$ y f^{IRe}	7	0.950	0.564	0.778	0.571	0.715	0.757

Tabla 4.12 Resultados del método propuesto y de sus principales ideas en el *corpus* *Plagiarised Short Answers*

²² El lector interesado en ver las tablas de los valores promedio y sus desviaciones estándar con las que se seleccionaron los valores de operación del método propuesto puede dirigirse al apéndice I.

4.9 Discusión de los resultados

Comparando la tabla 4.11 con la 4.8 y la tabla 4.12 con la 4.9, observamos que el método supera ligeramente los mejores casos de los resultados de los métodos de comparación; el incremento relativo de la eficacia del método propuesto es de entre 5% y 7% de la *precision global* y también se observa que, a diferencia del mejor método de referencia, el propuesto es más estable pues tuvo los mismos parámetros óptimos para ambos corpus, no se necesitó cambiar los parámetros como fue necesario con el MML o en con el tamaño del n-grama en los métodos de referencia.

En el *corpus* METER, al igual que en los métodos de referencia, la clase mejor evaluada es la que contiene plagio con un 0.862, que supera a la mejor de los métodos de referencia que tiene un 0.817. La clase de “*no plagio*” es inferior a la mejor de los métodos de referencia, en donde en una etapa intermedia del método propuesto se obtiene un 0.553 y los métodos de referencias alcanzan un 0.648.

En el *corpus* *Plagiarised Short Answers*, la clase “*no plagio*” es clasificada con una eficacia similar a la mejor de todos los experimentos de referencia que fue al usar el conjunto de n-gramas de {1-3}. Para la clase “*copias cercanas*” es el mejor método con una diferencia razonable, donde el mejor clasificador de los métodos de referencia para esta clase obtiene 0.684 y el método propuesto 0.778. En las clases con reescritura también se obtienen eficacias ligeramente superiores a todos los métodos de comparación; en “*revisión severa*”, la diferencia es de 0.558 a 0.564 y en “*revisión ligera*”, la diferencia es un poco más grande, de 0.465 a 0.571; esto es razonable, pues debe ser más difícil mejorar la eficacia en los casos de alta reescritura que en las de menor reescritura.

En el caso del desempeño global del método propuesto en el *corpus* METER, como se ha dicho, supera a los de referencia en la medida de *exactitud*, pero en el promedio del *f-measure* es ligeramente superado. El método propuesto tiene un comportamiento comparable a los mejores resultados de los métodos de referencia. Al utilizar las cadenas en este *corpus* se perdió eficacia en la clasificación de la clase “*no plagio*”; esto se puede

deber a que se haya forzado a los casos de no plagio a parecer a serlo, al tratar de encadenar secuencias que no fueran reutilizadas. Esto provoca también una pequeña pérdida en la medida global del promedio de la *f-measure*, de 0.704 a 0.694.

En el *corpus Plagiarised Short Answers* se obtuvo una mejora un poco más importante: el mejor resultado de referencia obtuvo una *exactitud* de 0.705 y el método propuesto 0.757, en el promedio de la *f-measure* se pasó de un 0.638 a un 0.715. En este *corpus* cada incorporación de las propuesta, del pesado y de las cadenas, ayudó a mejorar la efectividad; el pesado de distinción mejoró la clasificación en los caso de plagio, en especial, el de las “*copias cercanas*” y las cadenas mejora la detección de todas las clases, pero en mayor grado a las “*copias cercanas*” y a la “*revisión ligera*”.

Esta claro que la principal mejora sólo se dio en el *corpus Plagiarised Short Answers*, pues en el METER los resultados son sólo comparables, pero hay que recordar que justamente el *corpus Plagiarised Short Answers* es el verdadero corpus de DAP, pues el METER es sólo un *corpus* en el que se puede utilizar con tales fines. También tenemos que recordar que se ha comparado con los mejores resultados de los métodos de referencia, que se sabe que son buenos para esta tarea, y en especial, que la selección de los mejores resultados de los métodos de referencia se ha hecho *a posteriori*, después de completar varios experimentos en un amplio rango de parámetros, mientras que los parámetros del método propuesto fueron seleccionados con una heurística, la de G.I. de una preselección de atributos en una fracción del *corpus*. Y que los valores de los parámetros del método propuesto fueron más consistentes en ambos corpus que en los casos de los métodos de referencia, por lo que el método propuesto es menos sensible al cambio de *corpus*.

5 Conclusión

5.1 Recapitulación

En la investigación presentada se realizó una revisión del estado del arte de la identificación automática del plagio, y más concretamente, de la detección de plagio. Resultado de esta revisión se propuso una división con las jerarquías necesarias para organizar prácticamente todos los trabajos realizados en el área. De igual manera se organizó, utilizando parte de los lineamientos preexistentes, un ordenamiento de las tareas que engloban la identificación del plagio, entre las que se encuentran la detección del plagio; se sugeriría seguir esta división en los trabajos futuros, así como mantener la nomenclatura propuesta para las tareas particulares con el fin de mantener una organización de los posteriores desarrollos. La nomenclatura de las tareas particulares nace para atender la necesidad de plantear una normativa en donde se puedan determinar un marco de referencia de las problemáticas y

las necesidades a las cuales se deben enfocar cada una de las tareas, pues cada una de ellas permite (y requiere de) la especialización de las soluciones.

Respecto a los métodos reportados en el estado del arte, se obtuvo experimentalmente la confirmación del problema latente en la revisión de la literatura: algunos de los métodos, especialmente los que tienen buena eficacia, requieren de la elección de parámetros para su buen funcionamiento. Estos valores específicos son sensibles a la naturaleza del corpus y no se conoce forma de poder definirlo *a priori* a la evaluación del método.

5.2 Conclusiones

Finalmente, las conclusiones que se pueden extraer del trabajo realizado son:

- Tener varios atributos especializados en coincidencias textuales (o no) de varios tamaños aumenta la efectividad de la DAP; esto es demostrado por la mejora que presentaban los métodos de referencia que contaban varios atributos (el método del conjunto de atributos de cubrimientos con n-gramas expuesto por primera vez aquí como una extensión del encubrimiento de n-gramas) y al método propuesto.
- La utilización de cadenas, es decir, secuencias con cierta flexibilidad en el orden puede ayudar a la DAP, demostraron ser útiles en la DAP, especialmente en condiciones de plagio académico (como lo es el *corpus Plagiarised Short Answers*).
- Como es de sospecharse, las categorías o tipos de plagio más difíciles son aquellos en los que se involucra la reescritura. El método propuesto logra mejorar la discriminación del plagio que involucra reescritura, aunque este sigue siendo el escenario más difícil de tratar.
- Pesar la evidencia del plagio (el texto reutilizado) con base en la frecuencia que tienen los términos mejora la DAP.

- Los “métodos modernos de detección de plagio”, los aquí utilizados como métodos de referencia, son muy sensibles al cambio de parámetros y de la naturaleza del *corpus*.
- El método propuesto es, hasta cierto punto, insensible al cambio de la naturaleza del *corpus*.

5.3 Trabajo futuro

Después del trabajo realizado se abren varias posibles trabajos a desarrollar.

- Probar el método propuesto y las ideas en las que se basan en otros *corpus* de plagio. Esto se podría realizar con el *corpus* de la competencia internacional, PAN.
- Teniendo ya desarrollada la detección del plagio, es conveniente desarrollar el resto de las etapas para poder realizar la identificación de plagio automático; en especial la búsqueda de plagio, que ha adquirido una mayor atención en el último año debido a la creación de la competencia internacional. Para ello es necesario realizar una etapa anterior a la detección que asegure obtener un subconjunto de unas pocas fuentes, entre las cuales se encuentre la verdadera fuente del plagio, una etapa de IR podría ser suficiente.
- Hacer consideraciones especiales para los casos en los que el plagio está constituido primordialmente por cambios de palabras por sus sinónimos. Es cierto que en el presente trabajo se ha considerado que este problema es manejable, si se hace la suposición de que el cambio es una simple eliminación seguida de una inserción de una nueva palabra. Pero, independientemente de esta simplificación, resulta interesante analizar cuáles podrían ser las opciones para poder sustentar la identificación de las sustituciones que conserven el mismo sentido de aquellas que no lo hagan, sin requerir complejos recursos adicionales.

6. Referencias

- 1 Bouville, Mathieu: Plagiarism: Words and ideas, En: Science and Engineering Ethics, Volumen 14, Número 3, p.p. 311-322.
- 2 Barrón Cedeño, Alberto y Rosso, Paolo: On the relevance of search space reduction in automatic plagiarism detection, En: SEPLN: Procesamiento del lenguaje natural, N°. 43, 2009, p.p. 141-149.
- 3 Bilal Zaka: Theory and Applications of Similarity Detection Techniques, Tesis Doctoral: Institute for Information Systems and Computer Media, Graz University of Technology, febrero, 2009, p. 171.
- 4 ŘEHŮŘEK, Radim: Plagiarism Detection through Vector Space Models Applied to a Digital Library. En: Proceedings of Recent Advances in Slavonic Natural Language Processing, Volumen 1, Masaryk University, Brno 2008, p.p. 75-83.
- 5 Ceska, Zdenek: The Future of Copy Detection Techniques, En: Proceedings of the 1st Young Researchers Conference on Applied Sciences (YRCAS 2007), Pilsen, Czech Republic, November 2007, pp. 5-10.
- 6 Bouyssou, Denis, Martello, Silvano y Plastria, Frank: Plagiarism again: Sreenivas and Srinivas, with an update on Marcu, En: 4OR: A Quarterly Journal of Operations Research, Volumen 7, Número 1, Marzo 2009, p.p. 17-20.
- 7 Errami, Mounir, Hicks, Justin M., Fisher, Wayne, Trusty, David, Wren, Jonathan D., Long, Tara C. y Garner, Harold R.: Déjà vu: A database of highly similar citations in the scientific literature, En: Bioinformatics, Volumen 24, Número 2, Diciembre 2007, p.p. 243-249.
- 8 Iyengar R. N. y Mukunda N.: Notes on Plagiarism, En: SADHANA: Academy Proceedings in Engineering Science, India, 2008.

- 9 Monostori, Krisztián, Finkel, Raphael A., Zaslavsky Arkady B., Hodász, Gábor y Pataki, Máté: Comparison of Overlap Detection Techniques, En: Proceedings of the International Conference on Computational Science, Parte I, Amsterdam, abril, 21-24, 2002, p.p. 51-60.
- 10 Brin, Sergey, Davis, James y Garcia-molina, Hector: Copy Detection Mechanisms for Digital Documents, En: SIGMOD '95: Proceedings of the 1995 ACM SIGMOD International Conference on Management of Data, 22-25 Mayo 1995, p.p. 398-409.
- 11 Barrón-Cedeño, Alberto: Detección automática de plagio en texto, Tesis de maestría, Universidad Politécnica de Valencia, 2008. p. 75.
- 12 Clough, Paul, Gaizauskas, Robert, Piao, Scott y Wilks, Yorick: METER: MEasuring TEXT Reuse, En: Proceedings of the 40th Anniversary Meeting for the Association for Computational Linguistics (ACL-02), 7-12 Julio, Pennsylvania, Philadelphia, USA., 2002, p.p. 152-159.
- 13 Metzler, Donald, Bernstein, Yaniv, Croft, W. Bruce, Moffat, Alistair y Zobel, Justin: Similarity measures for tracking information flow, En: CIKM '05 Proceedings of the 14th ACM International Conference on Information and Knowledge Management, Bremen, Germany, 2005, p.p. 517-524.
- 14 Kang, NamOh, Gelbukh, Alexander y Han, SangYong: PPChecker: Plagiarism Pattern Checker in Document Copy Detection, En: Proceedings of the TSD-2006: Text, Speech and Dialogue, Lecture Notes in Computer Science, Volumen 4188, Brno, Czech Republic, Septiembre 11-15, 2006, p.p. 661-667.
- 15 Kang, NamOh y Han, SangYong: Document copy detection system based on plagiarism patterns, En: Proceedings Computational Linguistics and Intelligent Text Processing, 7th International Conference, CICLing 2006, Lecture Notes in Computer Science, Volumen 3878, Mexico City, Mexico, Febrero 19-25, 2006, p.p. 571-574.
- 16 Maurer, Hermann A., Kappe, Frank y Zaka, Bilal: Plagiarism - A Survey. En: Journal of Universal Computer Science, Volumen 12 Número 8, 2006, p.p. 1050-1084.

- 17 Lancaster, Thomas y Culwin, Fintan: A Visual Argument for Plagiarism Detection using Word Pairs, En: Proceedings of the Plagiarism Conference: Prevention, Practice and Policies 2004 Conference, Abril, 2004, p.p. 1-14.
- 18 Yang, Shen, Zhongshang, Yuan, Lu, Liu y Hui, Dong: Research of anti-plagiarism monitoring system model, En: Wuhan University Journal of Natural Sciences, Volumen 12, Número 5, septiembre de 2007, p.p. 937-940.
- 19 Vallés Balaguer, Enrique: Putting Ourselves in SME's Shoes: Automatic Detection of Plagiarism by the WCopyFind tool, En: Proceedings of the SEPLN'09 Workshop on Uncovering Plagiarism, Authorship and Social Software, Volumen 502, Donostia-San Sebastian, Spain, Septiembre 2009.2009, p.p. 34-35.
- 20 Culwin, Fintan y Lancaster, Thomas: Plagiarism, prevention, deterrence and detection, En: The Higher Education Academy, South Bank University, UK., Mayo 2001.
[http://www.heacademy.ac.uk/resources/detail/resource_database/id426_plagiarism_prevention_deterrence_detection revisado en 3 Junio].
- 21 Lancaster, Thomas y Culwin, Fintan: Visualising Intra-Corporal Plagiarism. En: Proceedings of 5th International Conference on Information Visualization (IV'01), London, July 2001, p.p. 289-296.
- 22 Gibson, Charles: A Primetime Investigation: Caught Cheating, En: Primetime Thursday (Television Broadcast), ABC Television Network, 29 April 2004. [revisado en 20/06/2010 <http://www.youtube.com/watch?v=hoH4yqVVp8c>]
- 23 Clough, Paul: Old and new challenges in automatic plagiarism detection. En: National UK Plagiarism Advisory Service, Volumen 76, Febrero, 2003, 391-407.
- 24 Clough, Paul: Plagiarism in natural and programming languages: an overview of current tools and technologies. En: Research Memoranda: CS-00-05, Department of Computer Science, University of Sheffield, UK, 2000.

- 25 Pataki, Máté: Distributed similarity and plagiarism search, En: Proceedings of the Automation and Applied Computer Science Workshop, Budapest, Hungary, 2006, p.p. 121-130.
- 26 Potthast, Martin, Stein, Benno, Eiselt, Andreas, Barrón-Cedeño, Alberto, and Rosso, Paolo. Overview of the 1st International Competition on Plagiarism Detection. En: SEPLN 2009 Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse (PAN 09), Septiembre 2009. p.p. 1-9.
- 27 Bernstein, Yaniv y Zobel, Justin: A Scalable System for Identifying Co-derivative Documents, En: String Processing and Information Retrieval, Lecture Notes in Computer Science, String Processing and Information Retrieval 11th International Conference, SPIRE 2004, Padova, Italy, Octubre 5-8, 2004, Volumen 3246, p.p. 1-11.
- 28 Basile, Chiara, Benedetto, Dario, Caglioti, Emanuele, Cristadoro, Giampaolo y Degli Esposti, Mirko: A Plagiarism Detection Procedure in Three Steps: Selection, Matches and "Squares", En: SEPLN 2009, 3rd PAN Workshop. Uncovering Plagiarism, Authorship and Social Software Misuse, Donostia-San Sebastian, Spain, Septiembre 2009, p.p. 19-23.
- 29 Grozea, Cristian, Gehl, Christian y Popescu, Marius: ENCOPLLOT: Pairwise Sequence Matching in Linear Time Applied to Plagiarism Detection, En: SEPLN 2009, 3rd PAN Workshop. Uncovering Plagiarism, Authorship and Social Software Misuse, Donostia-San Sebastian, Spain, Septiembre 2009, p.p.10-18.
- 30 Barrón-Cedeño, Alberto, Rosso, Paolo y Benedí, José-Miguel: Reducing the Plagiarism Detection Search Space on the Basis of the Kullback-Leibler Distance, En: Proceedings of the CICLing 2009, Computational Linguistics and Intelligent Text Processing, 10th International Conference, Lecture Notes in Computer Science, Volumen 5449, Mexico City, Mexico, Marzo 1-7, 2009, p.p. 523-534.
- 31 Stein, Benno, Meyer zu Eißén, Sven: Near Similarity Search and Plagiarism Analysis, En: From Data and Information Analysis to Knowledge Engineering, 29th

- Annual Conference of the German Classification Society (GfKI), Magdeburg, Marzo 9-11, 2005, p.p. 430-437.
- 32 Barrón-Cedeño, Alberto y Rosso, Paolo: On automatic plagiarism detection based on n-grams comparison, En: Proceedings of the Advances in Information Retrieval, 31th European Conference on IR Research, ECIR 2009, Toulouse, France, April 6-9, 2009. Lecture Notes in Computer Science, Volumen 5478, p.p. 696-700.
- 33 Grozea, Cristian y Popescu, Marius: Who's the Thief? Automatic Detection of the Direction of Plagiarism, En: Proceedings of the Computational Linguistics and Intelligent Text Processing, 11th International Conference, CICLing 2010, Lecture Notes in Computer Science, Volumen 6008, Iasi, Romania, Marzo 21-27, 2010, p.p. 700-710.
- 34 Kolcz, Aleksander, Chowdhury, Abdur y Alspector, Joshua: Improved robustness of signature-based near-replica detection via lexicon randomization, En: Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Seattle, Washington, USA, Agosto 22-25, 2004, p.p. 605-610.
- 35 Barrón-Cedeño, Alberto, Eiselt, Andreas y Rosso, Paolo: Monolingual Text Similarity Measures: A comparison of Models over Wikipedia Articles Revisions, En: Proceedings of ICON-2009: 7th International Conference on Natural Language Processing, Hyderabad, India, 2009, p.p. 29-38.
- 36 Zechner, Mario, Muhr, Markus, Kern, Roman y Granitzer, Michael: External and Intrinsic Plagiarism Detection Using Vector Space Models, En: SEPLN 2009, 3rd PAN Workshop. Uncovering Plagiarism, Authorship and Social Software Misuse, Donostia-San Sebastian, Spain, Septiembre 2009, p.p. 47-55.
- 37 Manning, Christopher D.: Introduction to Information Retrieval, Cambridge University Press, Cambridge, England, 1 Abril, 2009, p. 581.

- 38 Hoad, Timothy C. y Zobel, Justin: Methods for identifying versioned and plagiarized documents, En: Journal of the American Society for Information Science and Technology, Volumen 54, Número 3, Febrero, 2003, p.p. 203-215.
- 39 Shivakumar, Narayanan y Garcia-Molina, Hector: SCAM: A Copy Detection Mechanism for Digital Documents, En: Proceedings of 2nd International Conference in Theory and Practice of Digital Libraries (DL'95), Austin, Texas, June, 1995. p.p. 398-409.
- 40 Bao, Jun-Peng, Shen, Jun-Yi, Liu, Xiao-Dong, Liu, Hai-Yan y Zhang, Xiao-Di: Semantic Sequence Kin: A Method of Document Copy Detection, En: Proceedings of the Advances in Knowledge Discovery and Data Mining, Lecture Notes in Computer Science 3056, 8th Pacific-Asia Conference, PAKDD 2004, Sydney, Australia, Mayo 26-28, 2004, p.p. 529-538.
- 41 Si, Antonio, Leong, Hong Va y Lau, Rynson W. H.: CHECK: a document plagiarism detection system. En: Proceedings of the 1997 ACM symposium on Applied Computing, San Jose, CA, USA, Febrero 28 - Marzo 1, 1997, p.p. 70-77.
- 42 Chien-Ying, Chen, Yeh, Jen-Yuan y Ke, Hao-Ren: Plagiarism Detection using ROUGE and WordNet, En: Journal of Computing, Volumen 2 Número 3, Marzo 2010. p.p. 34-44.
- 43 Bao, Junpeng, Lyon, Caroline, Lane, Peter C. R., Ji, Wei y Malcolm, James A.: Comparing Different Text Similarity Methods, En: Technical Report 461, University of Hertfordshire, 2007.
- 44 Soledad Pera, Maria y Ng, Yiu-Kai: SimPaD: A Word-Similarity Sentence-Based Plagiarism Detection Tool on Web Documents. To appear in Web Intelligence and Agent Systems: An International Journal (JWIAS), 2010, IOS Press.
- 45 Seo, Jangwon y Croft, W. Bruce: Local text reuse detection. En: Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2008, Singapore, Julio 20-24, 2008. p.p. 571-578.

- 46 Guthrie, David, Allison, Ben, Liu, Wei, Guthrie, Louise and Wilks, Yorick: A Closer Look at Skip-Gram Modelling, En: Proceedings LREC'2006, Genoa, Italy, 2006, p.p. 1222-1225.
- 47 Alpert, Jesse y Hajaj y Nissan: Software Engineers, Web Search Infrastructure Team: Anunciado en: Blog oficial de Google en el 2008: We knew the web was big..., [revisado en 20/06/2010 <http://googleblog.blogspot.com/2008/07/we-knew-web-was-big.html>]
- 48 Lyon, Caroline, Malcolm, James y Dickerson, Bob: Detecting short passages of similar text in large document collections, En: Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA, June 2001, p.p. 118-125.
- 49 Lyon, Caroline, Barrett, Ruth y Malcolm, James: A theoretical basis to the automated detection of copying between texts, and its practical implementation in the Ferret plagiarism and collusion detector, En: Plagiarism: Prevention, Practice and Policies Conference, Newcastle, UK, June, 2004,
- 50 Barrón-Cedeño, Alberto y Paolo Rosso: Towards the exploitation of statistical language models for plagiarism detection with reference. En: Proceedings of the ECAI'08 PAN Workshop: Uncovering Plagiarism, Authorship and Social Software Misuse, Patras, Greece, 2008, p.p. 15-19.
- 51 Engels, Steve , Lakshmanan, Vivek y Craig Michelle: Plagiarism detection using feature-based neural networks, En: Proceedings of the 38th SIGCSE Technical Symposium on Computer Science Education, SIGCSE 2007, Bulletin archive Volumen 39 Número 1, Covington, Kentucky, USA, March 7-11, 2007, p.p. 34-38.
- 52 Wise, Michael J.: YAP3: IMPROVED DETECTION OF SIMILARITIES IN COMPUTER PROGRAM AND OTHER TEXTS. En: Proceedings of the 27th SIGCSE Technical Symposium on Computer Science Education, 1996, Philadelphia, Pennsylvania, USA, Febrero 15-17, 1996, p.p. 130-134.

- 53 Wise, Michael J.: Neweyes: A System for Comparing Biological Sequences Using the Running Karp-Rabin Greedy String-Tiling Algorithm, En: Proceedings of the Third International Conference on Intelligent Systems for Molecular Biology, Cambridge, United Kingdom, Julio 16-19, 1995, p.p. 393-410.
- 54 Broder , Andrei Z.: On the Resemblance and Containment of Documents, En: Compression and Complexity of Sequences (SEQUENCES'97), Salerno, Italy, 11-13 Junio 1997, p.p. 21-29.
- 55 Clough, Paul: Measuring text reuse, Tesis doctoral, University of Sheffield, UK., 2003
- 56 Witten, I., H., Frank, E.: Data Mining Practical Machine Learning Tools and Techniques, Elsevier, 2005.
- 57 Martin Potthast, Benno Stein, Andreas Eiselt, Alberto Barrón-Cedeño and Paolo Rosso: Overview of the 1st International Competition on Plagiarism Detection. En: SEPLN 2009, Workshop on Uncovering Plagiarism, Authorship and Social Software Misuse (PAN 09), Donostia-San Sebastian, Spain, September 2009. p.p. 1-9.
- 58 Gaizauskas, Robert, Foster, Jonathan, Wilks, Yorick, Arundel, John, Clough, Paul, Piao, Scott: The METER corpus: a corpus for analysing journalistic text reuse, En: Proceedings of Corpus Linguistics 2001, Lancaster, UK, p.p.214-223.
- 59 Clough, Paul. and Stevenson, Mark.: Creating A Corpus of Plagiarised Academic Texts. En: Proceedings of Corpus Linguistics Conference 2009, Liverpool. En prensa. Disponible en red [revisado en 20/06/2010 <<http://ir.shef.ac.uk/cloughie/papers/CL2009.pdf>>].

A. Apéndice I

En este apéndice se presentarán los experimentos y resultados con los que se seleccionaron todos los parámetros del método propuesto.

El método propuesto tiene un conjunto de constantes (c_2, c_3, c_4 y c_5) que deben de cumplir la inecuación AI.1

$$1 > c_2 > c_3 > c_4 > c_5 > 0 \quad (A.1)$$

Dentro del trabajo presentado en la tesis se propusieron 4 posibles funciones (A.2, A.3, A.4 y A.5) que podrían proporcionar los valores para las constantes. Es evidente que para que se cumpliera la ecuación AI.1 es necesaria una función decreciente; El motivo de la condición de la ecuación AI.1 es ir disminuyendo el valor o significancia que se le da a una palabra entre más lejano parece su uso en el documento sospechoso del uso que tuvo en la fuente. La incógnita es cuan severamente se debe de atenuar el valor conforme se aleja el uso de las palabras en ambos documentos, por esta razón las funciones propuestas tienen diferentes rapidez de decrecimiento.

$$c_n = \frac{1}{n} \quad (A.2)$$

$$c_n = \frac{1}{n^2} \quad (A.3)$$

$$c_n = \frac{1}{n^3} \quad (A.4)$$

$$c_n = \frac{1}{e^n} \quad (A.5)$$

En las tablas A.1 y A.2 se muestran los resultados de la G.I. del atributo f^{IRe} generado con cada uno de los esquemas de constantes de las funciones propuestas. En la medición del G.I. ambos *corpus* fueron divididos en 10 pliegues y los resultados mostrados son el promedio y la desviación estándar del I.G. obtenido en la variedad de pliegues.

Esquema de constantes por función	Promedio de G.I.	Desviación estándar
$\frac{1}{n}$	0.179	0.017
$\frac{1}{n^2}$	0.167	0.014
$\frac{1}{n^3}$	0.17	0.014
$\frac{1}{e^n}$	0.171	0.014

Tabla A.1 Muestra el promedio de la G.I. y su desviación estándar de f^{IRe} en las 10 particiones hechas en el *corpus* METER.

Esquema de constantes por función	Promedio de G.I.	Desviación estándar
$\frac{1}{n}$	0.915	0.107
$\frac{1}{n^2}$	0.863	0.11
$\frac{1}{n^3}$	0.863	0.08
$\frac{1}{e^n}$	0.863	0.08

Tabla A.2 Muestra el promedio de la G.I. y su desviación de f^{IRe} estándar en las 10 particiones hechas en el *corpus* *Plagiarised Short Answers*.

Como se aprecia en las tablas superiores, el mejor esquema de constantes para ambos *corpus* es la de la función menos decreciente, por lo tanto las constantes son:

$$c_2 = \frac{1}{2} \tag{A.5}$$

$$c_3 = \frac{1}{3} \tag{A.6}$$

$$c_4 = \frac{1}{4} \tag{A.7}$$

$$c_5 = \frac{1}{5} \tag{A.8}$$

Otro parámetro que se debe definir en el método propuesto es la longitud (en palabras) de la vecindad, v . Del mismo modo que los coeficientes, definimos el valor del parámetro v tomando la G.I. del atributo de f^{IRe} generado con cada valor particular de v . En las tablas A.3 y A.4 se muestran el promedio y desviación estándar de la G.I. con las que elegimos la v apropiada.

ν	Promedio de G.I.	Desviación estándar
3	0.172	0.017
5	0.179	0.017
11	0.164	0.019
15	0.172	0.017
21	0.175	0.019
25	0.174	0.011

Tabla A.3 Muestra el promedio de la G.I. y su desviación estándar de f^{IRe} en las 10 particiones hechas en el *corpus* METER para elegir el parámetro ν .

ν	Promedio de G.I.	Desviación estándar
3	0.915	0.107
5	0.915	0.107
11	0.850	0.088
15	0.850	0.080
21	0.836	0.088
25	0.834	0.077

Tabla A.4 Muestra el promedio de la G.I. y su desviación de f^{IRe} estándar en las 10 particiones hechas en el *corpus* *Plagiarised Short Answers* para elegir el parámetro ν .

Las tablas permiten ver que la ν que genera atributos que separan mejor las clases en el caso del *corpus* METER es claro que el mejor valor es ν igual con 5, en el caso del *corpus* *Plagiarised Short Answers*, la tabla indica que tanto el valor 3 y 5 generan los mejores atributos, es de preferir quedarse con la ν de mayor tamaño es decir 5 pero para estar seguro de la elección, realizamos un experimento para esclarecer si una ν de 5 es el mejor valor; para ello se en la tabla A.5 se muestra la detección de plagio con únicamente el atributo f^{IRe} generado con ambos valores de ν y como se esclarece en ese experimento el mejor valor para ν es 5.

ν	Promedio de f-measure.	Exactitud
3	0.602	0.684
5	0.613	0.694

Tabla A.5 Muestra el promedio de la f-measure y la exactitud al realizar la clasificación únicamente con el atributo f^{IRe} generado con los dos valores de ν que según la G.I. separan mejor las clases en el *corpus* *Plagiarised Short Answers*.

Otro parámetro del método que se debe de definir es el criterio con el que se considerar que existe una **tamaño de coincidencia grande** en el proceso de la obtención del índice de rescritura *IRe*, empleando la misma estrategia utilizada con lo parámetros

anteriores se calculo la G.I. sobre los 10 pliegues del corpus y también calculamos la desviación estándar de la G.I. en las diversas particiones. Los resultados se muestran en las tablas A.6 y A.7 en donde se encontró que para ambos corpus la mejor longitud (en palabras) para considerar una coincidencia grande fue de 3.

<i>Tamaño de coincidencia grande</i>	Promedio de G.I.	Desviación estándar
1	0.865	0.093
2	0.856	0.088
3	0.915	0.107
4	0.863	0.110
5	0.864	0.092

Tabla A.6 Muestra el promedio y la desviación estándar de la *ganancia de información, G.I.* del atributo f^{IRe} generados con ciertos de *tamaño de coincidencia grande*, sobre las 10 particiones del *corpus* METER.

<i>Tamaño de coincidencia grande</i>	Promedio de G.I.	Desviación estándar
1	0.054	0.003
2	0.056	0.006
3	0.057	0.006
4	0.056	0.008
5	0.055	0.007

Tabla A.7 Muestra el promedio y la desviación estándar de la *ganancia de información, G.I.* del atributo f^{IRe} generados con ciertos de *tamaño de coincidencia grande*, sobre las 10 particiones del *corpus* *Plagiarised Short Answers*.

El último parámetro a definir para el método propuesto es la m que define el número de atributos totales y el número de atributos dedicados a tamaños específicos. La estrategia es semejante a las utilizadas para definir los parámetros anteriores pero aquí el orden de los atributos importa, pues lo que se debe definir es el punto a partir del cual los atributos generados para una secuencia particular dejan de tener la capacidad de separar las clases y por tanto distinguir los casos de plagio. Las tablas A.6 y A.7 muestran la G.I. para los atributos específicos para las secuencias de longitud definida, n .

n	Promedio de G.I.	Desviación estándar
1	0.382	0.024
2	0.288	0.025
3	0.125	0.026
4	0.037	0.025
5	0.006	0.017
6	0	0

Tabla A.6 Muestra el promedio y la desviación estándar de la G.I. de f_n^{frg} en las 10 particiones hechas en el *corpus* METER para elegir el parámetro m .

n	Promedio de G.I.	Desviación estándar
1	0.552	0.075
2	0.244	0.025
3	0.287	0.062
4	0.031	0.030
5	0.019	0.058
6	0.345	0.046

Tabla A.7 Muestra el promedio y la desviación estándar de la G.I. de f_n^{frg} en las 10 particiones hechas en el *corpus* *Plagiarised Short Answers* para elegir el parámetro m .

En las tablas superiores vemos que en el caso del *corpus* METER el valor de la G.I. es decreciente, por lo que es evidente que sólo debemos definir el umbral a partir del cual cortaremos la representación y en la cual el resto de los atributos se aglutinaron en un único atributo. Decidimos que el umbral adecuado es de 0.1, por tanto la m es de 3. En el caso del *corpus Plagiarised Short Answers*, también tiene una m de 3; podría parecer que también vale la pena quedarse con el atributo de n igual a 6 pero eso implicaría incluir a los atributos de n igual con 4 y 5, lo que no es deseable pues estos atributos tienen poca capacidad de separar las clases, por lo que pueden sesgar el algoritmo de clasificación. La elección define los atributos de las secuencias que valen la pena tener desde n igual con 1 hasta donde los atributos tienen una G.I. menor a un umbral, ahora definido como 0.1.

B. Apéndice II

En este segundo apéndice se analiza a mayor detalle la eficacia del método de referencia comparado con los mejores métodos de referencia en el *corpus Plagiarised Short Answers*, el *corpus* que es realmente de plagio. Se realizaron experimentos sobre subconjuntos del *corpus* para evaluar la eficacia al separar los casos de no plagio y cada uno de los tipos de plagio y finalmente también se realizó un experimento para evaluar la eficacia al separar las clases de plagio. Los resultados se encuentran en la tablas B.1, B.2, B.3 y B.4, en el caso de la separación de los casos no plagiados y los caso de plagio de copias cercanas, el método propuesto es tan bueno como los mejores métodos de referencia; al separar el resto de los tipos de plagio de los casos de no plagio el método propuesto siempre supera a los métodos de referencia e inclusive la separación que realiza el método propuesto de los casos de plagio con revisión ligera de los casos de no plagio es perfecta.

Método	número de atributos	<i>f-measure</i>		Promedio de <i>f-measure</i>	<i>Exactitud</i>
		"no plagio"	"revisión severa"		
$f^{frag(\Psi^c)}$, $f^{dist(\Psi^c)}$ y f^{1Re}	7	0.961	0.919	0.940	0.947
{1-10}-gram	10	0.937	0.857	0.897	0.912
5-gram	1	0.937	0.857	0.897	0.912
SC (MML=8)	1	0.927	0.813	0.870	0.894

Tabla B.1 Comparación del método propuesto con los mejores métodos de referencia en la separación de las clases de "no plagio" y "revisión severa" del *corpus Plagiarised Short Answers*

Método	número de atributos	<i>f-measure</i>		Promedio de <i>f-measure</i>	<i>Exactitud</i>
		“no plagio”	“copias cercanas”		
$f^{frag(\psi^c)}, f^{dist(\psi^c)}$ y f^{lRe}	7	0.974	0.944	0.959	0.964
{1-10}-gram	10	0.974	0.944	0.959	0.964
5-gram	1	0.947	0.895	0.921	0.929
SC (MML=8)	1	0.974	0.944	0.959	0.964

Tabla B.2 Comparación del método propuesto con los mejores métodos de referencia en la separación de las clases de “no plagio” y “copias cercanas” del corpus *Plagiarised Short Answers*

Método	número de atributos	<i>f-measure</i>		Promedio de <i>f-measure</i>	<i>Exactitud</i>
		“no plagio”	“revisión ligera”		
$f^{frag(\psi^c)}, f^{dist(\psi^c)}$ y f^{lRe}	7	1	1	1	1
{1-10}-gram	10	0.987	0.974	0.980	0.982
5-gram	1	0.960	0.923	0.941	0.947
SC (MML=8)	1	0.987	0.973	0.980	0.982

Tabla B.3 Comparación del método propuesto con los mejores métodos de referencia en la separación de las clases de “no plagio” y “revisión ligera” del corpus *Plagiarised Short Answers*

Método	número de atributos	<i>f-measure</i>			Promedio de <i>f-measure</i>	<i>Exactitud</i>
		“revisión severa”	“copias cercanas”	“revisión ligera”		

$f^{frag(\psi^c)}$, $f^{dist(\psi^c)}$ y f^{IRe}	7	0.621	0.757	0.483	0.621	0.631
{1-10}-gram	10	0.653	0.611	0.483	0.582	0.596
5-gram	1	0.545	0.632	0.37	0.515	0.508
SC (MML=8)	1	0.45	0.632	0.37	0.484	0.456

Tabla B.4 Comparación del método propuesto con los mejores métodos de referencia en la eficacia de la separación de las 3 tipos de plagio (“*revisión severa*”, “*copias cercanas*” y “*revisión ligera*”) que están clasificados en el *corpus Plagiarised Short Answers*.

C. Apéndice III

Los resultados del método propuesto expuestos en el capítulo 4 de evaluación son obtenidos con los parámetros definidos con las estrategias de selección guiadas con la ganancia de información (ver apéndice I). Estas estrategias de selección de los valores de los parámetros permiten obtener una eficacia superior a los mejores resultados de los métodos de referencia, pero no garantizan obtener los mejores resultados posibles del método propuesto. En la tabla C.1 se expone la eficacia del método propuesto con los valores obtenidos mediante la estrategia de G.I. y otros desempeños que superaban la eficacia reportada en el capítulo 4 al utilizar parámetros distintos a los obtenidos a partir de la estrategia empleada.

<i>v</i>	<i>m</i>	Tamaño de coincidencia grande	número de atributos	<i>f-measure</i>				Promedio de <i>f-measure</i>	<i>Exactitud</i>
				"no plagio"	"revisión severa"	"copias cercanas"	"revisión ligera"		
5	3	3	7	0.950	0.564	0.778	0.571	0.715	0.757
5	6	3	13	0.937	0.609	0.811	0.5	0.714	0.768
5	3	2	7	0.962	0.651	0.833	0.5	0.736	0.789

Tabla C.1 Se exponen la eficacia del método propuesto con los valores sugeridos por los métodos de selección de parámetros expuestos en el apéndice I en la primera fila, y en las filas sucesivas otras evaluaciones que resultaron ser mejores al utilizar valores distintos a los indicados por la G.I. en el corpus *Plagiarised Short Answers*.