Detección Intrínseca de Plagio

Javier Pérez Afonso

Dirigido por:

Paolo Rosso
Universitat Politècnica de València

J. Fernando Sánchez-Vega
Instituto Nacional de Astrofísica, Óptica y Electrónica
México

UNIVERSITAT
POLITÈCNICA
DE VALÈNCIA

Trabajo Final de Máster desarrollado dentro del Máster en Inteligencia Artificial, Reconocimiento de Formas e Imagen Digital

Valencia, Febrero de 2013

Resumen

Debido a la gran cantidad de documentos susceptibles de plagio, la tarea de la detección de plagio es imposible de realizar de forma manual. Por este motivo, desde hace algunos años se están desarrollando métodos automatizados que ayuden al experto lingüistica a detectar los casos de plagio.

Existen dos grandes enfoques en la detección de plagio, la detección externa y la detección intrínseca. La detección externa consiste en comparar los fragmentos de un documento sospechoso con un conjunto de documentos para encontrar los fragmentos plagiados. La detección intrínseca usa únicamente las características internas del documento, para indentificar los fragmentos plagiados debido a que no se dispone de un conjunto de documentos de referencia. En esta tesis nos centramos en la detección intrínseca mediante la aplicación de diversas medidas estilísticas tradicionales y de dos nuevas medidas propuestas a las que llamamos Índice de Concordancia Local o ICL y, su generalización, Índice General de Concordancia Local o IGCL.

Para realizar la detección proponemos dos métodos que emplean el conjunto de medidas estilísticas presentadas a lo largo de la tesis. El primer método, es capaz de aprender de un conjunto de documentos ejemplo las características necesarias, para poder realizar la separación de los documentos con plagio de aquellos que se encuentran libres de plagio. El segundo, permite reconocer los fragmentos de los documentos que no conservan el estilo de la mayoría del documento y que por tanto podrían considerarse como fragmentos sospechosos de plagio, para el cual usamos algoritmos de clustering.

Los métodos aquí desarrollados, fueron evaluados empleando estándares fijados por la competencia internacional de detección de plagio (PAN). Los resultados obtenidos, muestran que la eficacia de la nueva medida estilística propuesta es similar al empleo del conjunto de las medidas estilísticas tradicionales, sin requerir el uso de un gran conjunto simultáneo de características para detectar el plagio.

Índice general

Parte I

Introducción

Introducción

1 Introducción

La acción de plagiar según la Real Academia de la Lengua Española [19], es:

Copiar en lo sustancial obras ajenas, dándolas como propias.

Existe un amplio conjunto de objetos susceptibles de ser plagiados: las imágenes, la música, las ideas y los documentos. En el presente trabajo nos vamos a centrar en el plagio escrito, es decir, el plagio de documentos. En este ámbito el acto de plagiar, significa incorporar fragmentos de un documento escrito por otro autor sin darle el crédito correspondiente.

En los últimos años, los casos de plagio de documentos se han visto incrementados. Desde la aparición de Internet se ha desarrollado otro fenómeno que ha suscitado gran interés, que se considera como plagio basado en la reutilización de contenidos.

Existen varios estudios sobre el incremento de casos de plagio en el ámbito universitario, en particular en [12] se muestra que desde el 1999 hasta el 2005 hubo un incremento del 30 % de los estudiantes que había confesado haber cometido plagio. Otro estudio realizado por la Universidad de Stanford publicó que se produjo un aumento del 126 % entre 1998 y 2001 [12].

El problema del plagio es casi imposible de solucionar de forma manual, debido a la gran cantidad de documentos existentes tanto en bibliotecas digitales como en Internet. Debido a esto, se ha incrementado la investigación en este campo y se están desarrollando sistemas automatizados que nos permiten realizar esta tarea de manera más fácil y en un tiempo razonable.

Hoy en día, en la red se pueden encontrar múltiples herramientas para la detección de plagio en documentos. Una de ellas es *WriteCheck* [1] la cual sirve para prevenir el plagio en los trabajos, aunque también se puede utilizar para detectar fragmentos o documentos plagiados. Esta aplicación muestra el porcentaje de secciones qué pueden estar escritas de manera similar en otros documentos. En España, hay empresas que comienzan a involucrarse en la detección de plagio, Citilab[2] ha desarrollado una herramienta gratuita[3] que se puede utilizar para detectar casos de plagio. En este caso esta aplicación nos muestra las fuentes de las que se ha cometido el plagio.

En las siguientes secciones de este capitulo introduciremos la problemática del plagio, la descripción del problema que aborda este trabajo, su motivación y sus objetivos y para finalizar describiremos la estructura de esta tesis.

1.1 Problemática del plagio

El problema del plagio de documentos en el ámbito universitario, laboral y científico se ha incrementado notablemente en los últimos años, principalmente debido al fácil acceso a documentos electrónicos.

Se pueden distinguir varios tipos de plagio, según Iyer y Singh [15] podemos destacar tres:

- Palabra por palabra. Se basa en la copia exacta de fragmentos de un documento sin incluir su fuente.
- De referencias. Se da cuando una referencia está en un documento y se incluye en otro documento sin haber leído el origen.
- De autoría. Ocurre cuando un autor dice ser creador de un trabajo que fue realizado por otro.

Según esta clasificación, los tipos de plagios que son susceptibles de ser detectados mediante el análisis de documentos son el plagio palabra por palabra y el de autoría. El de referencias es casi imposible, ya que si se incluye la cita y la referencia correctamente no se podría distinguir si es un caso de plagio o no.

El enfoque que proponemos en este trabajo aborda los tipos de plagio, nombrados anteriormente, ya que se analizan los cambios de estilo entre el autor del documento y los fragmentos plagiados para intentar detectar esos fragmentos plagiados.

[1] https://www.writecheck.com/static/features.html
[2] https://www.citilab.eu
[3] https://copionic.citilab.eu/

1.2 Descripción del problema

La tarea de la detección de plagio consiste en identificar si un documento contiene fragmentos plagiados. En este ámbito existen dos enfoques principales, la *detección externa de plagio* y la *detección intrínseca de plagio* [2].

La idea general de la *detección externa de plagio*, es encontrar los casos de plagio mediante la comparación del documento sospechoso con las posibles fuentes de los fragmentos que fueron plagiados. Por otra parte, cuando no se tiene ningún corpus de referencia (conjunto de documentos) en el cual buscar el (o los) fragmentos plagiados, nos debemos centrar en las características propias del documentos para intentar detectar si existe algún indicio de plagio. A este último caso, enfocado en las características internas del documento, se llama *detección intrínseca de plagio.*

En el ámbito de la *detección intrínseca de plagio*, las medidas estilísticas nos pueden ayudar a representar los fragmentos de un documento y analizar los cambios de estilo para detectar los fragmentos plagiados de un documento.

1.3 Motivación

Debido al incremento de casos y sus repercusiones en los diferentes ámbitos afectados (universitarios, científicos, laborales, etc.) desde hace algunos años se han investigado múltiples enfoques en la detección de plagio para la creación de sistemas automáticos.

La misma tecnología que ayuda a realizar el acto de plagiar de manera sencilla, nos puede ayudar a detectarlo. Algunos de los métodos que se utilizan en la búsqueda de información, tratamiento de documentos y traducción estadística pueden ser usados y adaptados para la detección de plagio [1].

En la realidad, existen situaciones en las que no disponemos de un conjunto de documentos originales para comprobar la existencia de plagio en un documento sospechoso, por lo que no podemos recurrir a la *detección externa de plagio.* Estas situaciones pueden ocurrir por diversos motivos, por ejemplo: algunas de las bibliotecas digitales son de acceso restringido, las posibles fuentes de donde se ha realizado el plagio no siempre estarán digitalizadas. Por estos motivos, en este trabajo exploramos la *detección intrínseca de plagio* ya que es útil para realizar la detección sin la comprobación de la fuente original.

Aunque las medidas estilísticas han demostrado ser eficientes para caracterizar el estilo de un documento y han tenido una amplia aceptación por parte de la comunidad lingüistica, no han sido aún exploradas en detalle para la realización de la tarea de la detección intrínseca de plagio. Es fácil observar que estas medidas pueden proporcionar la información necesaria para distinguir entre los documen-

tos plagiados y no plagiados. Razón por la cual proponemos usar las medidas estilítisticas para la detección intrínseca de plagio.

La necesidad de indentificar los fragmentos de un documento con un estilo similar en la detección intrínseca de plagio, motiva al empleo de los algoritmos de agrupamiento (*clustering*). Estos algoritmos permiten la creación de grupos de fragmentos con características semejantes, que facilitan la tarea de identificar las anomalías estilísticas, propias del plagio, a un experto lingüista.

1.4 Objetivos

Los objetivos de esta investigación se han dividido en generales y particulares, y estos se describen a continuación:

1. Generales
    - Investigar la aplicación de las medidas estilísticas para la *detección intrínseca de plagio*.
2. Particulares
    - Obtener la representación de los fragmentos del documento mediante medidas estilísticas.
    - Generar agrupaciones mediante algoritmos conocidos para identificar los fragmentos sospechosos de plagio de un documento.

1.5 Estructura de la tesis

A continuación, se describen los demás capítulos de los que consta este trabajo:

- En la sección 2, abordaremos el estado del arte de la detección de plagio, donde describiremos los dos principales enfoques, la *detección externa de plagio* y la *detección intrínseca de plagio*. El segundo enfoque se desarrolla con más detalle y abordaremos las medidas estilísticas dentro de éste, ya que es el enfoque en el que nos hemos centrado en esta investigación.
- En la sección 3, describiremos los métodos desarrollados para la tarea de la *detección intrínseca de plagio*: una nueva medida a la que denominamos *Índice de concordancia local* y su generalización para caracterizar a todo el documento sospechoso. También presentamos en esta sección dos métodos, el supervisado donde abordaremos como vamos a analizar el comportamiento de las medidas estílisticas y por último, el método no supervisado donde describiremos el uso de las medidas estilísticas mediante el agrupamiento de fragmentos con similitudes estilísticas.

- En la sección 4, describiremos el corpus utilizado para realizar los experimentos de evaluación de los métodos desarrollados y analizaremos con los resultados que hemos obtenido.
- En la sección 5, incluimos las conclusiones a las que hemos podido llegar tras el trabajo realizado, así como el camino futuro que planteamos seguir tras esta investigación.

Parte II

Detección de Plagio

Detección de Plagio

2 Detección de plagio

El plagio desde el punto de vista del procesamiento del lenguaje natural se modela como una reutilización de texto [11]. La reutilización de textos consiste en el uso del texto de una fuente previamente publicada. Dentro de la reutilización de textos distinguimos varios tipos: la copia exacta, la reescritura añadiendo o quitando palabras, el resumen donde se puede realizar una modificación de oraciones o una síntesis del contenido de un texto y la traducción de textos. Estos cuatro tipos están organizados de acuerdo al grado de dificultad, comenzando por los más sencillos de detectar.

Cuando el texto reutilizado no está correctamente citado, dando cuenta del origen, se considera plagio. La detección de plagio es la tarea que trata de identificar los fragmentos que han sido reutilizados. Cuando hablamos de la tarea de *detección de plagio*, no podemos olvidarnos de la tarea hermana, la de atribución de autoría [25, 21, 8], que consiste en intentar determinar cual es el autor de un documento comparando con otros textos. La tarea de la atribución de autoría no sólo se basa en analizar documentos completos para comprobar si están escritos por un único autor, sino que además, intenta analizar un texto o fragmentos del mismo para intentar verificar si realmente están escritos por el autor que se atribuye el mérito.

Dentro de la *detección de plagio* existen dos enfoques principales, el primero, descrito en la sección 2.1, es la *detección externa de plagio* que se basa en la comparación de documentos originales con el documento sospechoso, para detectar los fragmentos con indicios de que están copiados o reutilizados. El segundo enfoque, descrito en la sección 2.2, en el que se centra este trabajo es la *detección intrínseca de plagio*, la cual se basa en detectar si un documento contiene fragmentos plagiados sin utilizar fuentes externas (es decir, no se puede verificar cual es la fuente original de dicho documento).

Para poder realizar la detección de plagio, es crucial seleccionar un conjunto de características del documento que nos permitan discriminar los documentos sospechosos de estar plagiados de los originales. En [24, 1], se proponen una serie de características que se pueden utilizar para detectar posibles casos de plagio [4].

- Vocabulario utilizado. Analizar el vocabulario utilizado en algún escrito, con respecto a documentos escritos previamente por el mismo estudiante. La existencia de una alta cantidad de vocabulario nuevo, podría ayudar a determinar si un estudiante realmente escribió el documento o no.
- Cambios de vocabulario. Cambios de vocabulario significativos a lo largo de todo el documento. A diferencia de la anterior, esta característica analiza el cambio interno del uso del vocabulario y no respecto a algunos ejemplos de escritura del presunto autor.
- Texto incoherente. Si un texto fluye de manera inconsistente o confusa podría deberse a la existencia de secciones plagiadas.
- Puntuación. Es muy poco probable que dos autores utilicen los signos de puntuación exactamente de la misma manera.
- Cantidad de texto común entre documentos. Es poco usual que dos documentos escritos de manera independiente compartan grandes cantidades de texto.
- Errores en común. Resulta altamente improbable que dos textos independientes tengan los mismos errores de escritura.
- Distribución de las palabras. Es poco frecuente que la distribución en el uso de las palabras a través de textos escritos independientemente sea la misma.
- Estructura sintáctica del texto. Un indicador de plagio es que dos textos compartan una estructura sintáctica común.
- Largas secuencias de texto en común. Es poco probable que dos textos independientes (incluso cuando traten el mismo tema), compartan largas secuencias de caracteres o palabras consecutivas.
- Orden de similitud entre textos. Si existe un conjunto significativo de palabras o frases comunes en dos textos, y si el orden de ocurrencia de las coincidencias de los textos es el mismo, puede haber un caso de plagio.
- Dependencia entre ciertas palabras y frases. Un autor tiene preferencias sobre el uso de ciertas palabras y frases. Encontrarlas en un trabajo realizado por otro, debe ser considerado sospechoso.

[4] Estas caracterísiticas se pueden utlizar en cualquier ámbito, aunque estén especificadas para el académico

- Frecuencia de palabras. Es poco común que las palabras halladas en dos textos independientes sean usadas con la misma frecuencia.
- Preferencia por el uso de sentencias cortas o largas. Los autores pueden tener una marcada preferencia sobre la longitud de las sentencias. Dicha longitud podría ser poco usual en comparación de otras características.
- Legibilidad del texto. Resulta improbable que dos autores compartan las mismas métricas de legibilidad.
- Referencias incongruentes. La aparición de referencias en el texto que no se encuentran en la bibliografía o viceversa, son disparadores de un posible caso de plagio.

Hay algunas de estas características que se pueden poner en duda a la hora de ser utilizadas en la detección de plagio. Un ejemplo de esto, es la característica correspondiente al vocabulario utilizado, ya que una persona que está en un proceso de formación académica puede haber incrementado su vocabulario. Otras características como la cantidad de texto en común y largas secuencias en común, añaden dificultad a la tarea de la detección de plagio, en concreto por el tiempo de cómputo que se requiere a la hora de realizar las comparaciones exhaustivas necesarias, entre documentos sospechosos de ser plagiados y los originales.

Estas características nombradas anteriormente, junto con otras, se han utilizado en diferentes enfoques en la *detección intrínseca de plagio*. Estos enfoques serán explicadas a lo largo de la sección 2.2.

2.1 Detección externa de plagio

Para resolver el problema de la *detección de plagio*, la primera solución intuitiva que se realiza es buscar en otros documentos para detectar si el documento tiene fragmentos plagiados. Esta es la idea básica en la que se basa la aproximación llamada: *detección externa de plagio*. Cuando se han encontrado fragmentos posiblemente plagiados, son verficados por un experto lingüista. Uno de los grandes problemas del método basado en la búsqueda del material plagiado es el tiempo de ejecución, que es demasiado grande. De ahí la importancia de abordar el problema de la detección de plagio de manera automática.

En la actualidad, existen multitud de formas de realizar la comparación de textos sospechosos con otros originales para la *detección externa de plagio*. La mayoría de los métodos realizan la detección externa de plagio siguiendo tres etapas generales: descomposición, comparación e identificación. En la descomposición, los documentos se fragmentan en sub-cadenas (n-gramas) de caracteres. En la comparación se usan las sub-cadenas de los documentos para comparar las secciones

de los documentos sospechosos con los originales, identificando el número de subcadenas iguales. Por último, la identificación, establece un criterio para detectar las secciones plagiadas, normalmente se emplea un umbral que establece la cantidad de material común necesario para determinar el plagio. En [5, 17, 7] los autores usan el método de las 3 etapas explicado anteriormente, en [28], se utiliza una descomposición empleando fragmentos de siete palabras (7-gramas de palabras). En [13], se propone un método que emplea varios ciclos de las tres etapas variando el tamaño de los fragmentos de la etapa de descomposición para disminuir el número de posibles fuentes de plagio en cada una de las sucesivas comparaciones.

En esta sección se ha descrito de forma general *la detección externa de plagio* ya que no es el enfoque princial de esta tesis. En siguiente sección se tratará con más detalle la *detección intrínseca de plagio* basada en la caracterización del estilo de los fragmentos de un documento para indentificar los fragmentos plagiados.

2.2 Detección intrínseca de plagio

Actualmente, con la cantidad de documentos que existen, la detección de plagio es una tarea casi imposible de realizar y en ocasiones no disponemos de un conjunto de documentos con el cual comparar, para buscar el posible material plagiado. Por esta razón, en los últimos años, se han desarrollado métodos para detectar el plagio sin consultar otros documentos, a la cual se le llama *detección intrínseca de plagio*. En la *detección intrínseca de plagio*, se analizan las características propias del documento, ya que no disponemos de un conjunto de documentos con los que comparar.

Según Stein y Meyer zu Eissen, en [4], la tarea de la detección intrínseca de plagio se puede definir de la siguiente forma:

Dado un documento d, escrito por un autor, se requiere identificar los fragmentos en d que se derivan de otro autor. El análisis intrínseco del plagio es como un problema de clasificación de una sola clase.

El problema que nos podemos encontrar en esta clasificación, según los autores nombrados anteriormente, es que sólo tenemos información de una única clase. En el ámbito de la *detección intrínseca de plagio*, los documentos o partes del documento, forman parte de esa única clase y los documentos de otro autor constituyen la clase atípica, por lo tanto, sólo tenemos un documento como única fuente para identificar los cambios de estilo.

Los principios básicos de la *detección intrínseca de plagio* son [16, 10]:

- Cada autor tiene un estilo diferente de escritura.
- Coherencia en todo el texto del estilo de escritura del autor.

- Las características de un estilo son difíciles de imitar y manipular.

Aunque la tarea de *detección intrínseca de plagio* es relativamente nueva, la cuantificación de las características de estilo de la escritura ha sido investigada por varios autores desde los años 40 [14, 26, 3, 27, 6]. En este ámbito, se han definido una serie de medidas para evualuar las características de estilo en un documento. Estas medidas se llaman *medidas estilísticas*. Las principales medidas estilísticas son:

1. *Índice Gunning Fox* [27]. Esta medida se ha desarrollado para medir el grado de legibilidad que tiene un documento escrito. El resultado que se obtiene, nos índica los años aproximados que necesita una persona en formación para comprender un texto:
$$I_G = 0{,}4 \left(\frac{|palabras|}{|sentencias|} + 100 \frac{|palabrascomplejas|}{|palabras|} \right) \quad (1)$$
   - *palabras complejas* : son las que tienen al menos tres sílabas (menos nombres propios y sufijos, como es, ed o ing).
   - $|palabras|$: número de palabras en el texto evaluado.
   - $|sentencias|$: número de sentencias en el texto evaluado.
2. *Índice Flesch-Kincaid* [26]. Esta medida está relacionada con la anterior, ya que mide de igual forma los años de educación requeridos para la comprensión de un documento:
$$I_{FK} = 206{,}835 - 1{,}015\lambda - 84{,}6\beta \quad (2)$$
   - λ : número de palabras dividido entre el número de sentencias.
   - β: número de sílabas dividido entre el número de palabras.
3. *Función R* [3]. Esta medida obtiene la variedad de vocabulario de un documento escrito:
$$R = \frac{100log(M)}{M^2} \quad (3)$$
   - M : es el número de palabras en el documento.
4. *Función K* [14]. Esta función es una variación de la función R, que mide el cálculo de la riqueza del vocabulario:
$$K = \frac{10^4(\sum_{i=1}^{\infty} i^2 V_i - M)}{M^2} \quad (4)$$
   - M : es el número de palabras en el documento.
   - V_i : es el vector que representa el número de veces que aparece cada palabra en el documento.
5. *Longitud media de las Palabras* [6]. Esta medida, es el promedio de la longitud de las palabras calculada en caracteres.
6. *Longitud media de las Sentencias* [6]. Esta medida, es el promedio de la longitud de las sentencias del documento calculada en palabras.

Para detectar si un documento o fragmento de un documento está plagiado, no es suficiente con calcular dichas características para todo el documento. Existen una gran variedad de formas en las que se miden las diferencias entre los fragmentos. En [28], [9], [20], se han utilizado medidas de disimilitud calculadas sobre bolsas de n-gramas de los fragmentos que se están comparando. En [18], las palabras de interés son los sustantivos, verbos, adverbios, palabras comunes, *hapax legomenon* [5] y las palabras cortas. En [22], la atención se centra en las palabras vacías y los sufijos de las palabras y finalmente, en [28], se utilizan toda una categoría de expresiones que llaman "marcadores discursivos".

En el trabajo que presentamos, se propone medir la similitud entre los fragmentos mediante la comparación de las medidas estilísticas en lugar de comparar elementos del texto. Al realizar la comparación, utilizando las medidas estilísticas que han sido introducidas por la comunidad de expertos lingüistas, como métodos de identificación y caracterización del estilo de escritura. Será más sencillo para un experto, por ejemplo un lingüistia forense, poder interpretar los resultados y exponerlos con mayor claridad ante un tribunal.

[5] es una palabra que sólo se produce una vez dentro de un contexto, ya sea en las obras de un autor o en sólo texto.

Parte III

Métodos Desarrollados

Métodos Desarrollados

3 Métodos desarrollados

En esta sección, abordaremos los trabajos que hemos realizado en esta tesis de máster en la tarea de la *detección íntrinseca de plagio*. Los métodos se basan en el estudio de los cambios de estilo, lo que se conoce como estilometría. La estilometría analiza las características internas, para detectar las posibles diferencias de estilo entre el autor del documento que se está analizando.

En este trabajo, se usan algunas de las medidas estilísticas que se han descrito en el capítulo anterior, debido a que captan distintas características del estilo de los autores de un documento. De igual modo, presentamos una nueva medida basada en el método de detección propuesto por Stamatatos [9] a la que llamamos *Índice de concordancia local*, donde se realiza la comparación entre los fragmentos del documento y se analiza la ruptura del estilo en contextos cercanos. Con ello se detectan los fragmentos que son sospechosos de estar plagiados. El índice de concordancia local está descrito en la sección 3.2 y en la sección 3.3, hablaremos de la generalización de esta medida llamada *Índice general de concordancia local* donde se caracteriza un documento en su globalidad.

Las diferencias de estilo capturadas por las medidas estilísticas, son analizadas empleando un método supervisado y uno no supervisado. En el análisis supervisado, se investiga la eficiencia de la discriminación de los fragmentos plagiados, usando un conjunto de entrenamiento para calcular la variación normal estilística de los documentos que presentaremos en la sección 3.3. Dentro de este primer método se utiliza un *detector intrínseco de plagio*. Este detector está basado en la expectativa de identificar la existencia (o inexistencia) de fragmentos plagiados mediante el grado de variación de la medida estilística. Por otro parte, en el análisis no supervisado investigamos la aplicación de algoritmos de *agrupamiento* (o *clustering*) que nos deberían permitir obtener los conjuntos de los fragmentos que poseen similitudes estilísticas. Por tanto son fragmentos pertenecientes a un mismo autor así como se presentará con más detalle en la sección 3.4.

3.1 Medidas estilísticas

Las medidas estilísticas se pueden agrupar según las características de estilo capturadas en un documento; por una parte, están las medidas que nos dan información sobre la legibilidad del documento, el *índice Gunning Fox* e *índice Flesch-Kincaid*; por otro lado, tenemos las medidas que obtienen información sobre la riqueza de vocabulario, *función K* y la *función R*; por último, están las medidas que nos miden el promedio de la longitud de las palabras y de las sentencias [6].

Para la realización de este trabajo se utiliza la implementación de las medidas disponibles en *Stylysis* [7]. Esta aplicación, que está disponible online, permite la visualización gráfica del resultado de estas medidas, lo cual es una valiosa ayuda para observar su comportamiento.

3.2 Índice de concordancia local

Como parte del estudio que aquí presentamos, hemos propuesto y evaluado el comportamiento de una nueva medida estilística, la cual llamamos *Índice de concordancia local (ICL)*, usada para la caracterización estilística de un fragmento de texto.

El propósito del *Índice de concordancia local* es caracterizar un fragmento de texto según la concordancia que tenga con el contexto en el que se encuentra. Este índice, permite plasmar los cambios bruscos que el autor realiza (e inclusive la suavidad con la que se presentan estos cambios) desde el punto de vista temático (cuando cambia o introduce nuevos temas en un texto) así como desde el punto de vista estilístico (cuando introduce algunas secciones en un tono distinto al resto del documento).

Para poder realizar la caracterización del *Índice de concordancia local* de un fragmento de texto, es necesario comparar dicho fragmento con los fragmentos que se encuentran a su alrededor. Realizamos la comparación del fragmento caracterizado con los fragmentos de su contexto, usando la medida de disimilitud basada en la comparación de las frecuencias de trigramas de caracteres definida en la ecuación (6) y usada por Stamatatos en su detector intrínseco de plagio.

En la ecuación (6), los fragmentos de texto f_A y f_B, son fragmentos que se encuentran en el mismo contexto. Hemos definido diferentes niveles de contexto según la distancia de los fragmentos comparados por la ecuación (6). Por tanto, el *Índice de concordancia local* de nivel k, es la comparación del fragmento caracterizado con el fragmento del texto que se encuentra a una distancia de **k** fragmentos, tal como se expresa en la ecuación (6).

Teniendo un documento sospechoso conformado por la sucesión de N fragmentos:

$$d_S = \{f_1, f_2, ..., f_i, f_{i+1}, ..., f_{N-1}, f_N\} \quad (5)$$

[6] Estas medidas han sido descritas en el capítulo dos de esta tesis, concretamente en la sección 2.2.

[7] http://memex2.dsic.upv.es:8080/StylisticAnalysis/es/index.jsp

El *Índice de concordancia local* de nivel k para el fragmento i es:

$$ICL_k(f_i) = \Sigma_{g \epsilon f_i} C_{cl} \frac{\left(\frac{F_{f_i}(g) - F_{f_{i+k}}(g)}{F_{f_i}(g) + F_{f_{i+k}}(g)}\right)^2}{|f_i|} \tag{6}$$

- $F_{f_i}(g), F_{f_{i+k}}(g)$: frecuencias del n-gramas de caracteres g en los fragmentos de texto f_i, f_{i+k}.
- $|f_i|$: Número de caracteres en el fragmento f_i.
- C_{cl} : Constante de concordancia local, igual a 1000.

3.3 Índice general de concordancia local

En esta sección, introducimos una nueva medida denominada *Índice general de concordancia local* o *IGCL*, que utilizamos para caracterizar un documento en su globalidad, y nos es útil en el análisis supervisado, debido a que con esta medida podemos comparar características globales de varios documentos, al contrario que en el método no supervisado en el que comparamos las características de los fragmentos del mismo documento.

Para calcular el *IGCL* para un documento sospechoso d_S, se calcula el *Índice de concordancia local* con nivel de 1 a N para todos los fragmentos del documento sospechoso y posteriormente se obtiene el promedio de cada variación estándar de los diferentes niveles del *Índice de concordancia local*, como lo indica la ecuación (7).

$$IGCL(d_S) = \frac{1}{N} \sum_{i=1..N} \sigma(\bigcup_{j=1..N} \{ICL_j(f_i)\}) \tag{7}$$

- $\cup_{j=1..N} ICL_j(f_i)$ es el conjunto las valoraciones del *Índice de concordancia local* a nivel j de todos los fragmentos del documento sospechoso d_S.
- $\sigma(C)$ es la desviación estándar del conjunto de valores del ICL.

El *IGCL* indica el nivel de concordancia global del documento. Para la *detección intrínseca de plagio,* esta medida es de gran utilidad ya que en un documento que contiene fragmentos plagiados, la concordancia interna del documento será menor debido a la diferencia de estilo (y algunos casos también pequeñas diferencias temáticas) de los fragmentos que han sido plagiados.

En las siguientes secciones presentamos dos métodos de detección intrínseca de plagio. El primero es un método supervisado basado en el uso de las medidas estilísticas para discriminar documentos plagiados evaluando el comportamiento de esas medidas dentro de unos parámetros de normalidad. El segundo es un método no supervisado que usa agrupamientos de fragmentos similares de un documento mediante la cohesión de las medidas estilísticas.

3.4 Método supervisado

En el método supervisado evaluamos la eficiencia de la discriminación de los fragmentos plagiados. Este método puede utilizar cualquier medida de las mencionadas en apartados anteriores, como atributo para caracterizar al documento. Cada una de estas medidas captura una característica del estilo diferente de los fragmentos que componen el documento. Debido a esto, se pueden utilizar para discriminar los fragmentos sospechosos comparando la variación de estas medidas.

Este análisis se ha realizado en dos etapas, una de aprendizaje y la segunda de clasificación. En la etapa de aprendizaje, se realiza el cálculo de las variaciones estilísticas de cada una de las medidas usadas en el método. Estas variaciones de las medidas se han calculado con el promedio de la desviación estándar de un conjunto de documentos de entrenamiento. En la etapa de clasificación, en el análisis de discriminación realizado, se ha evaluado la precisión del *detector intrínseco de plagio* basado en la simple aplicación de un umbral en cada una de la desviaciones estándar de las medidas estilísticas del documento.

El *detector intrínseco de plagio* propuesto está basado en la expectativa de identificar la existencia (o inexistencia) de fragmentos plagiados mediante el grado de variación de la medida estilística. El umbral se calcula con un conjunto de documentos.

3.5 Método no supervisado

Así como se ha descrito anteriormente, la tarea de *detección intrínseca de plagio* consiste en determinar los fragmentos sospechosos de plagio de un documento. nosotros proponemos en este método realizar por medio de la cohesión de las características de los fragmentos de dicho documento.

La naturaleza de esta tarea involucra la identificación de grupos de fragmentos de texto similares, por tanto es sumamente útil usar herramientas de *agrupamiento* o *clustering*, ya que nos permiten agrupar distintos fragmentos por medio de sus distintas características. Se han usado las diferentes medidas explicadas en las secciones anteriores, que capturan diferentes aspectos del estilo, para obtener una agrupación que integre toda la complejidad de los diferentes aspectos estilísticos.

El método propuesto, está basado en la representación de un documento sospechoso conformado por su conjunto de fragmentos *f*, así como se presenta en la ecuación (8):

$$d_S = \{f_1, f_2, ..., f_i, f_{i+1}, ..., f_{N-1}, f_N\} \tag{8}$$

Donde cada documento, posteriormente se representa por medio de una matriz, a la que denominamos D^M, conformada por el conjunto de los atributos de cada fragmento calculados mediante la función A, la cual calcula el conjunto de atributos de cada fragmento f del documento d_S. Formalmente se puede expresar de la siguiente manera (9):

$$D^M = < \bigcup_{i}^{|D|} A(f_i) > \tag{9}$$

Posteriormente, se realiza la función $cluster_q$ sobre la matriz D^M, para obtener la agrupación de los fragmentos de textos:

$$cluster_q(D^M) = \{G_1, G_2, ..., G_q\}, \tag{10}$$

- $G_1, G_2, ...G_q$ son los grupos de fragmentos obtenidos por la apliación del algoritmo de clustering la matriz D^M

Los clusters obtenidos tienen que ser dividos en dos grupos: los fragmentos plagiados y los no plagiados. Para obtener el conjunto de los fragmentos plagiados, primero se ordenan los q grupos de mayor tamaño a menor tamaño, tomando el tamaño como el número de fragmentos que lo componen. Los primeros grupos en este ordenamiento contendrán los fragmentos que comparten un estilo mayoritario en el documento, pues es la mayor cantidad de fragmentos similares entre sí, y por tanto deben ser considerados como grupos o fragmentos no plagiados. Por otro lado, los últimos grupos en este ordenamiento contendrán los fragmentos anomálos del texto, pues son distintos a la mayoría de los fragmentos, considerándolos fragmentos plagiados.

Para determinar el número de grupos pertenecientes a la clase no plagiados (NP) y la clase plagiados (P), se recorre el ordenamiento de los clusters partiendo de los grupos con mayor tamaño hasta los grupos más pequeños. A los primeros clusters se le asigna la clase NP y se va calculando el porcentaje de los fragmentos del documentos pertenecientes a esta clase. Cuando se haya cubierto un determinado porcentaje del documento, al resto de grupos no cubiertos se le asignará la clase P. Finalmente, los pasajes encontrados como plagiados, son la unión de aquellos fragmentos contiguos pertenecientes a un mismo grupo declarado como plagiado. Se mantiene la separación entre los grupos declarados como plagiados, para facilitar al experto lingüista la visualización y reconocimiento de los grupos de anomalías detectas por el método.

Parte IV

Experimentos y Resultados

Experimentos y resultados

4 Experimentos y resultados

Se ha realizado el análisis de aplicabilidad de las medidas estilísticas para la evaluación de los distintos métodos descritos en la sección 3, para detectar los fragmentos plagiados. En la siguiente subsección de este capítulo, en la sección 4.1, describiremos el corpus que ha sido usado para la evaluación de nuestros métodos. En la sección 4.2, introduciremos las métricas de evaluación utilizadas para medir la eficacia de los métodos propuestos, para abordar la tarea de la *detección intrínseca de plagio*, y por último, en la sección 4.3, analizaremos los resultados obtenidos después de aplicar los métodos propuestos en esta tesis.

4.1 Corpus

El corpus PAN-PC-11[8] empleado en nuestros experimentos ha sido creado para la competición de detección de plagio del Lab del CLEF sobre Uncovering Plagiarism, Authorship, and Social Software Misuse (PAN). En esta competición se puso a disposición de los usuarios interesados un recurso para desarrollar un sistema y evaluar el corpus en tarea la *detección intrínseca de plagio*. El corpus contiene 5000 documentos, donde existen documentos plagiados y otros libres de plagio, que pueden tener más de un caso de plagio con longitudes diferentes. El corpus desarrollado para esta competición, se encuentra en texto plano codificado en UTF-8. Los documentos se encuentran separados en carpetas, donde cada carpeta contiene un total de 500 documentos. Esta colección de documentos está acompañada de unos ficheros con extensión XML, en los cuales se realiza una descripción de los ficheros que contenían plagio, el tipo de plagio[9] que contenían los documentos sospechosos y la longitud de los fragmentos plagiados. La distribución de los documentos que contienen fragmentos plagiados y de documentos no plagiados se muestra en la siguiente tabla 1.

A continuación, describimos un subconjunto de prueba de 174 documentos, seleccionados aleatoriamente del corpus estándar de evaluación. Utilizamos este subconjunto

[8] http://www.uni-weimar.de/cms/medien/webis/research/corpora/corpus-pan-pc-11.html
[9] Los tipos de plagio puede ser artificial y simulado

Documentos	Nº de Doc
No Plagiados	2766
Plagiados	2234
Total	**5000**

Tabla 1: División de los documentos del corpus PAN

para analizar la eficacia de cada medida estilística por separado empleando el método supervisado. La distribución de documentos que contienen fragmentos plagiados y de documentos libres de plagio en el subconjunto se muestra en la siguiente tabla:

Documentos	Nº de Doc
No Plagiados	54
Plagiados	120
Total	**174**

Tabla 2: División del subconjunto de documentos

Este subconjunto de documentos se ha utilizado para el cómputo del método supervisado y en él definimos tres conjuntos de particular interés: *(**P**lagiados)*, que contiene todos los documentos de esta subparte del corpus que tienen al menos un fragmento plagiado, *(**No P**lagiados)* son los documentos que no contienen ningún plagio y finalmente nombramos como *(Todos)*, a la unión de los dos conjuntos *(**P**lagiados)* y *(**No P**lagiados)*. La definición de estos conjuntos nos servirá para la descripción de los conjuntos de entrenamiento en el método supervisado.

4.2 Métrica de evaluación

En esta sección, describiremos las medidas de evaluación empleadas en los análisis de los métodos supervisados y no supervisados. Una de las medidas de evaluación que utilizamos a nivel de documento se llama *precisión a nivel de documento* o $precisión_D$ y solamente es utilizada en el método supervisado. Esta medida la utilizamos en la evaluación del *detector intrínseco de plagio* basado en la simple aplicación de un umbral en cada una de la desviaciones estándar de las medidas estilísticas del documento. Como umbral se ha usado el promedio de la desviación estándar de la medida estilística calculada en todo el conjunto de prueba (U_{Todos}) y calculado exclusivamente sobre documentos no plagiados ($U_{\mathbf{No\ P}lagiados}$).

El método no supervisado se ha evaluado con las medidas ocupadas en el PAN y definidas en el artículo [23]. Estas medidas son, *recall* (11), *precision* (12) y *F-measure* (13) [10]

[10] En esta competición se usa una medida llamada *granularidad*, la cual no empleamos ya que el método que proponemos usa agrupamientos de los fragmentos del texto por medio de la cohesión estilística, y parte de la motivación de este trabajo es permitir al experto lingüista observar y

Para todas las medidas se ha utilizado la siguiente notación:

- s hace referencia a un fragmento plagiado de un conjunto S de todos los fragmentos plagiados.
- r denota la detección del R de todas las detecciones; S_R nos indica el subconjunto de S para los que existen detecciones en R.
- $|s|$,$|r|$ hace referencia a la longitud en caracteres de s y r.
- $|S|$,$|R|$ representan la longitud de dichos conjuntos.

Estas medidas siguen las siguientes fórmulas:

$$recall = \frac{1}{|S|} \sum_{s \epsilon S} \frac{\alpha(s_i)}{|s_i|} \tag{11}$$

donde $\alpha(s_i)$ representa el número de caracteres detectados de s_i.

$$precision = \frac{1}{|R|} \sum_{r \epsilon R} \frac{\beta(r_i)}{|r_i|} \tag{12}$$

donde $\beta(s_i)$ representa el número de caracteres plagiados de s_i donde R_s representa el número de detecciones de s en R.

$$F\text{-}measure = 2 * (\frac{recall * precision}{recall + precision}) \tag{13}$$

4.3 Resultados

4.3.1 Resultados del método supervisado

En esta sección, se muestran los resultados del método supervisado. Hemos evaluado el potencial para discriminar los documentos con fragmentos plagiados de los documentos **N**o **P**lagiados con las diferentes medidas estilísticas.

La desviación estándar de las métricas índica la variación que tiene el estilo (observado por una medida en particular) a lo largo del documento. La introducción de fragmentos plagiados en un documento deben de modificar la dinámica de la variación del estilo en un documento. En la tabla 3, presentamos el promedio de las desviaciones estándar de cada una de las medidas estilísticas calculadas sobre el subconjunto completo de prueba [11].

comparar los diferentes grupos encontrados como plagiados por el algoritmo de clustering sin un postproceso de unificación.

[11] Las características de este subconjunto de prueba utilizado se encuentran en la sección 2.2

Medidas	Conjunto de documentos		
	No Plagiados	Plagiados	Todos
IGCL	3.12	2.73	2.84
Gunning Fox	13.69	**15.47**	14.69
Longitud Sentencias	28.66	**31.67**	29.33
Longitud Palabras	0.55	0.51	0.52
Función R	798.55	739.62	765.48
Índice Flesch-Kincaid	0.88	**0.90**	0.89
Función K	49.19	**49.77**	49.50

Tabla 3: Promedio de las desviaciones estándar de las medidas estilísticas de cada subconjunto de documentos

Se espera que el promedio de la desviación estándar de los documentos con fragmentos plagiados (*Plagiados*), sea superior que cuando se calcula sobre todos los documentos (*Todos*), debido a que los los fragmentos plagiados tendrán valores estilísticos distintos al del material original del documento, lo que provocará una mayor dispersión de la población de las mediciones estilísticas aumentando la desviación estándar.

En los resultados experimentales mostrados en la tabla 3, se observa que la hipótesis anterior se cumple para las medidas estilísticas de Gunning Fox, longitud de Sentencias, Índice Flesch-Kincaid y la función K. Estas mismas medidas estilísticas cumplen con la contra parte de la misma hipótesis, que los documentos libre de plagio tienen una desviación estándar menor.

La tabla 4 muestra los resultados de los *detector intrínseco de plagio* con los conjuntos de documentos, el conformado por la unión del conjunto de los documentos con fragmentos plagiados (**P**lagiados) y el conjunto de los documentos (**N**o **P**lagiados), a la que nos referimos como U_{todos} y exclusivamente el conjunto formado por los documentos con fragmentos libres de plagio, a que llamamos $U_{libredeplagio}$, en los que se calculó el umbral (U).

Medidas	Precisión	
	U_{todos}	$U_{libredeplagio}$
IGCL	79.89 %	86.21 %
Gunning Fox	64.94 %	61.49 %
Longitud Sentencias	28.16 %	27.01 %
Longitud Palabras	86.21 %	81.61 %
Función R	93.10 %	78.16 %
Índice Flesch-Kincaid	54.02 %	51.72 %
Función K	62.64 %	61.49 %

Tabla 4: Precisión del *Naive-detector* con cada umbral

De los resultados de la tabla 4, se ha encontrado que el umbral calculado mediante todos los documentos, obtiene mejores resultados que el que se calcula sólo con documentos libres de plagio excepto en la medida *IGCL*, que hemos propuesto en este trabajo, en la cual el umbral calculado sólo con documento libre de plagio obtiene mejores resultados. Tener un método de detección entrenado únicamente con documentos libres, como es el caso del *detector intrínseco de plagio* con *IGCL* es más conveniente pues no se requiere de tener ejemplos positivos los cuales son más dificiles de obtener en un escenario realista. En la tabla (2) también es necesario resaltar que las funciones que presentan mejor potencial

para detectar el plagio son la Función R, la longitud de las Palabras y la *IGCL*, siendo esta última una propuesta presentada en este trabajo.

A continuación, presentamos el análisis experimental de algunas de las características que presentan las diferentes medidas estilísticas consideradas en este estudio. En la figura (1) se presenta la comparativa del comportamiento que presenta la medida propuesta, *IGCL*, variando el tamaño del los documentos, para los documentos libres de plagio y con plagio.

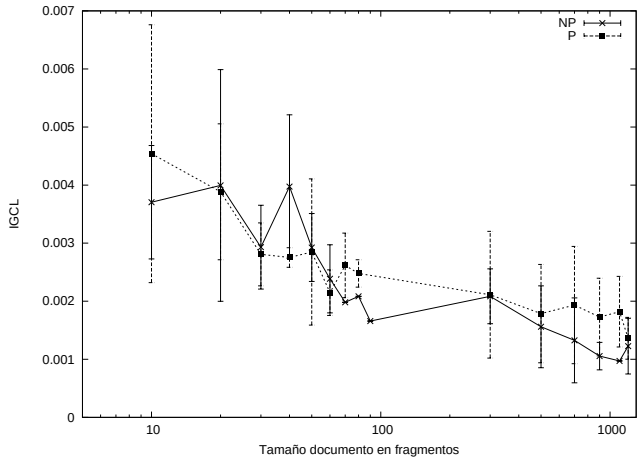

Figura 1: Resultados del IGCL por tamaño de documentos donde P: documentos con fragmentos Plagiados; NP: documentos No Plagiados.

En la figura 1 se logra apreciar como la medida *IGCL* se va estabilizando, obteniendo una menor desviación estándar, para documentos más grandes. También se obtiene valores de concordancia mayores en los documentos más extensos. Este comportamiento se explica debido a que en los documentos cortos se tiene que resumir un contenido que en muchas ocasiones puede abordar más de un tema lo que provoca que exista grandes saltos temáticos y estilísticos, debidos al limitado espacio.

Un comportamiento interesante que se observa en la figura 1 es la existencia de una mayor desviación estándar en los documentos con plagio que en los documentos libres de plagio independientemente de su longitud. Lo cual indica que la existencia de plagio introduce elementos con menos concordancia con el documento, como es de esperar al existir fragmentos plagiados.

En las figuras 2, 3, 4, 5, 6 y 7, se muestran los histogramas separados para los documentos no plagiados y para los documentos con plagio de los valores de desviación estándar de las diferentes medidas estilísticas estudiadas. Al comparar el comportamiento de la desviación estándar de los documentos con plagio frente a los documentos no plagiados observamos que el indice Gunning Fox y la medida de la Longitud de las Palabras son las que presentan un comportamiento significativamente distinto para cada conjunto de documentos, por otro lado las medidas del función R y del índice Flesch no

tienen diferencias claras en su comportamiento entre los dos conjuntos de documentos aunque la pequeña diferencia en la función R ha sido suficiente para posicionarse como la mejor medida para detectar casos en el análisis realizado con el *detector intrínseco de plagio*.

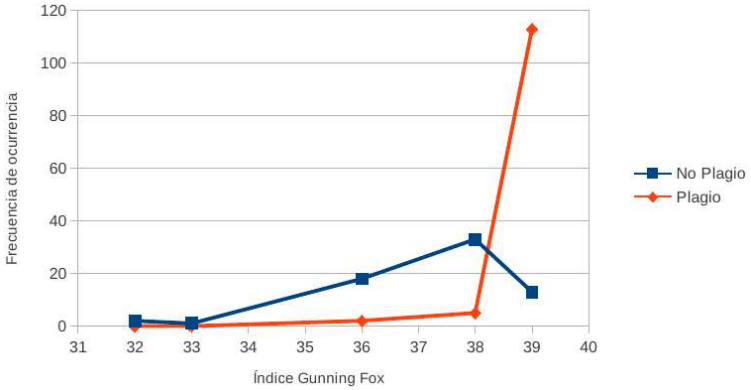

Figura 2: Histograma del Índice Gunning Fox

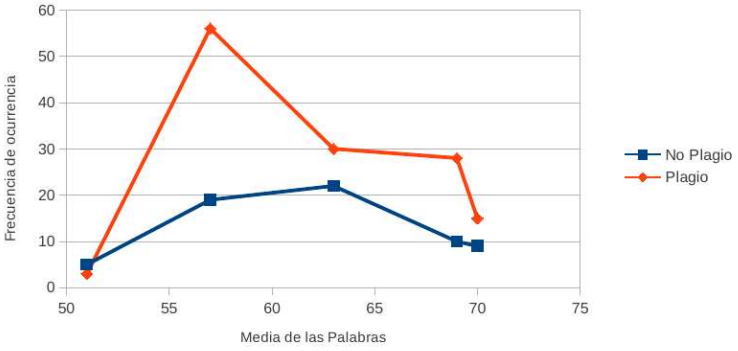

Figura 3: Histograma de la medida Longitud media de las Palabras

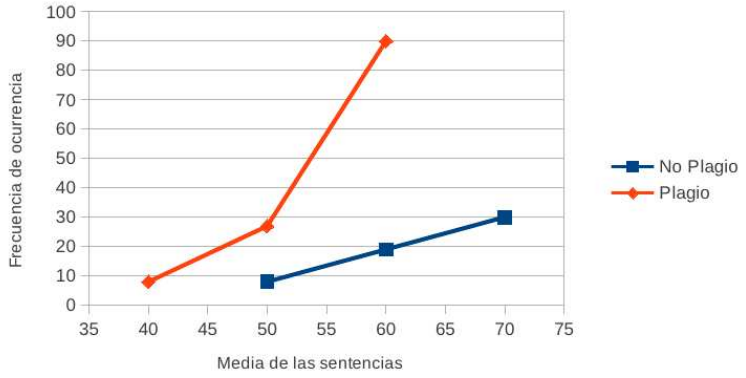

Figura 4: Histograma de la Longitud media de las Sentencias

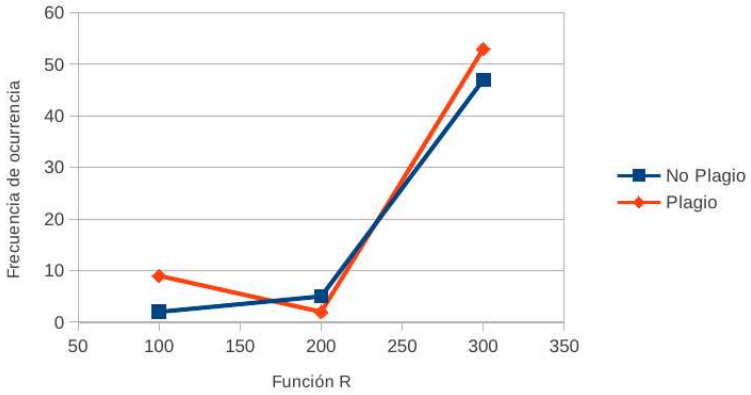

Figura 5: Histograma con la función R

4.3.2 Resultados del método no supervisado

En esta subsección, presentamos la evaluación del método no supervisado. Los experimentos se han realizado utilizando el Cluto, una herramienta desarrollada por George Karypis[12] un profesor que pertenece al Departamento de Ciencias de la Computación e Ingeniería de la Universidad de Minnesota. Esta herramienta sirve para realizar agrupaciones de documentos analizando las características comunes existentes entre éstos. Su aplicación no fue directa, y fue necesario hacer consideraciones especiales para la detección, tal como lo explicamos en la sección 3.5, donde describimos el método no supervisado.

Hemos evaluado diferentes algoritmos de agrupación para determinar cual es el más apropiado para la *detección intrínseca de plagio*. Los algoritmos usados son *RB*, *AGGLO*,

[12]http://glaros.dtc.umn.edu/gkhome/

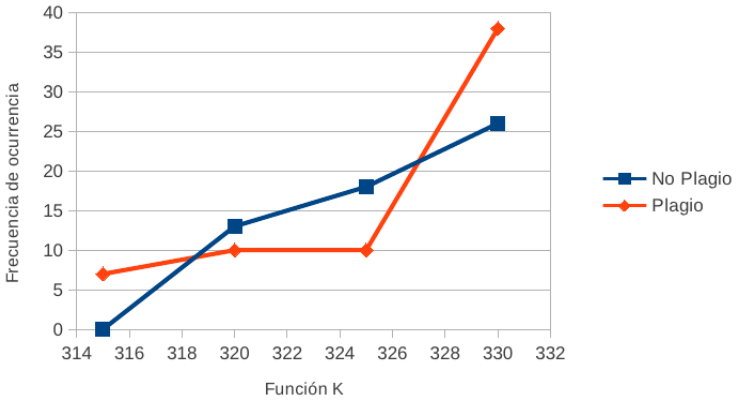

Figura 6: Histograma con función K

Figura 7: Histograma del Índice Flesch

DIRECT y *GRAPH*. En el *RB*, la solución del cluster se calcula mediante la realización de una secuencia de *k - 1* bisecciones repetidas. En este algoritmo, la primera matriz se agrupa en dos, uno de estos grupos se selecciona y se bisecciona otra vez más. Este proceso continua hasta el número deseado de cluster.

AGGLO, calcula la solución de los *k-caminos* utilizando el paradigma de la aglomeración, cuyo objetivo es optimizar localmente (minimizar o maximizar) una agrupación particular usando una función como criterio. La solución se obtiene parando el proceso de aglomeración cuando k clusters lo permiten.

En el *DIRECT*, la solución de los cluster se calcula simultáneamente buscando todos los grupos k. En general, el cálculo de un agrupamiento k-vías directa es más lenta que la agrupación a través de bisecciones repetidas. En términos de calidad, para valores

razonablemente pequeños de k (por lo general menos de 10-20), el enfoque directo conduce a mejores grupos que los obtenidos a través de bisecciones repetidas. Sin embargo, a medida que aumenta k, el enfoque repetición de la bisectriz tiende a ser mejor que la agrupación directa.

Finalmente, el *GRAPH*, calcula la solución del cluster modelando primero los objetos utilizando una gráfica de vecinos más cercanos (cada objeto se convierte en un vértice, y cada objeto está conectado a sus otros objetos similares), y luego divide el gráfico en k-grupos utilizando una partición del gráfico llamada min-cut. El gráfico si tiene más de un componente conectado, el algoritmo regresa a la solución (k+m)-vías, donde m es el número componentes conectadas en el gráfico.

Con esta herramienta se han realizado dos tipos de pruebas. La primera consiste en realizar con Cluto dos clustering, que asignarán a dos clases, NP (**N**o **P**lagiado) y P (**P**lagiado) que se basa en ver el cluster que mayor número de fragmentos tiene y asignarle la clase NP, mientras que el cluster con menor número de fragmentos se le asigna la clase P. La segunda consiste en hacer lo mismo pero con un número variable de clusters, donde contaremos los clusters con mayor número de fragmentos hasta llegar al menos al 50 % del número total de fragmentos del documento y ese conjunto de clusters le asignaremos la clase NP y al resto la clase P. Estas pruebas se han realizado paras todas las medidas estilísticas, ICL y la unión de las dos medidas.

Para verificar la eficiencia de los métodos descritos en las secciones anteriores se ha utilizado el evaluador[13] de la competición PAN. Esta eficiencia se mide según unas medidas predeterminadas, que son *recall* según la ecuación (11), *precisión* siguiendo la ecuación (12) y la *F-measure* calculada por medio de la ecuación (13). A continuación, se muestran los resultados obtenidos en la siguientes tablas 5, 6, 7 y la 8, donde *NC* es el Número de Cluster, *ME* son todas Medidas Estilísticas y *ME&ICL* es una unión de las dos medidas *ME* e *ICL*.

[13]Script facilitado en la competición PAN para evaluar la eficiencia los métodos desarrollados.

NC	Medidas	Recall	Precision	F-Measure
2	ICL	0,8829	0,0917	0,1661
	ME	0,8829	0,0917	0,1661
	ME&ICL	0,8745	0,0918	0,1662
5	ICL	0,9446	0,0901	0,1646
	ME	0,9408	0,0928	0,1689
	ME&ICL	0,9369	0,0927	0,1687
10	ICL	0,9590	0,0868	0,1593
	ME	0,9608	0,0918	0,1677
	ME&ICL	0,9565	0,0921	0,1680
15	ICL	0,9608	0,0853	0,1566
	ME	0,9650	0,0906	0,1657
	ME&ICL	0,9597	0,0904	0,1652
20	ICL	0,9605	0,0846	0,1555
	ME	0,9662	0,0899	0,1645
	ME&ICL	0,9621	0,0901	0,1647
30	ICL	0,9584	0,0841	0,1547
	ME	0,9668	0,0887	0,1624
	ME&ICL	0,9624	0,0890	0,1629
50	ICL	0,9603	0,0839	0,1543
	ME	0,9674	0,0876	0,1607
	ME&ICL	0,9637	0,0876	0,1607
70	ICL	0,9601	0,0839	0,1543
	ME	0,9684	0,0867	0,1591
	ME&ICL	0,9644	0,0870	0,1595
100	ICL	0,9591	0,0838	0,1542
	ME	0,9686	0,0863	0,1584
	ME&ICL	0,9642	0,0862	0,1582

Tabla 5: Resultados del método no supervisado utilizando las medidas ICL, ME, la unión de estas ME&ICL y diferentes números de cluster (NC) empleando el algoritmo *RB*

NC	Medidas	Recall	Precision	F-Measure
2	ICL	0,5274	0,0995	0,1674
	ME	0,5274	0,0995	0,1674
	ME&ICL	0,5192	0,0987	0,1658
5	ICL	0,7507	0,0815	0,1470
	ME	0,9011	0,0904	0,1643
	ME&ICL	0,8987	0,0941	0,1704
10	ICL	0,9131	0,0842	0,1541
	ME	0,9508	0,0921	0,1679
	ME&ICL	0,9485	0,0925	0,1685
15	ICL	0,9193	0,0826	0,1515
	ME	0,9539	0,0911	0,1663
	ME&ICL	0,9367	0,0906	0,1652
20	ICL	0,9090	0,0820	0,1504
	ME	0,9505	0,0904	0,1651
	ME&ICL	0,9173	0,0899	0,1638
30	ICL	0,8820	0,0823	0,1505
	ME	0,9442	0,0884	0,1617
	ME&ICL	0,8865	0,0887	0,1612
50	ICL	0,8051	0,0832	0,1508
	ME	0,9284	0,0877	0,1602
	ME&ICL	0,8080	0,0873	0,1576
70	ICL	0,7342	0,0834	0,1497
	ME	0,9128	0,0869	0,1586
	ME&ICL	0,7352	0,0866	0,1549
100	ICL	0,7056	0,0833	0,1490
	ME	0,9073	0,0859	0,1570
	ME&ICL	0,7066	0,0855	0,1526

Tabla 6: Resultados del método no supervisado utilizando las medidas ICL, ME, la unión de estas ME&ICL y diferentes números de cluster (NC) empleando el algoritmo *AGGLO*

NC	Medidas	Recall	Precision	F-Measure
2	ICL	0,8829	0,0917	0,1661
	ME	0,8829	0,0917	0,1661
	ME&ICL	0,8745	0,0918	0,1662
5	ICL	0,9476	0,0905	0,1652
	ME	0,9437	0,0945	0,1718
	ME&ICL	0,9406	0,0942	0,1713
10	ICL	0,9573	0,0868	0,1592
	ME	0,9596	0,0935	0,1704
	ME&ICL	0,9565	0,0934	0,1701
15	ICL	0,9372	0,0853	0,1564
	ME	0,9584	0,0917	0,1675
	ME&ICL	0,9403	0,0919	0,1674
20	ICL	0,9273	0,0846	0,1550
	ME	0,9579	0,0910	0,1662
	ME&ICL	0,9203	0,0909	0,1654
30	ICL	0,8930	0,0841	0,1538
	ME	0,9509	0,0896	0,1638
	ME&ICL	0,8956	0,0895	0,1627
50	ICL	0,8125	0,0834	0,1513
	ME	0,9379	0,0878	0,1606
	ME&ICL	0,8141	0,0877	0,1583
70	ICL	0,7347	0,0835	0,1499
	ME	0,9277	0,0872	0,1594
	ME&ICL	0,7387	0,0867	0,1552
100	ICL	0,7065	0,0834	0,1492
	ME	0,9229	0,0863	0,1579
	ME&ICL	0,7076	0,0861	0,1535

Tabla 7: Resultados del método no supervisado utilizando las medidas ICL, ME, la unión de estas ME&ICL y diferentes números de cluster (NC) empleando el algoritmo *DIRECT*

NC	Medidas	Recall	Precision	F-Measure
2	ICL	0,9057	0,0948	0,1716
	ME	0,9057	0,0948	0,1716
	ME&ICL	0,8993	0,0953	0,1724
5	ICL	0,2167	0,1030	0,1396
	ME	0,6847	0,0985	0,1723
	ME&ICL	0,6822	0,0982	0,1717
10	ICL	0,2010	0,1024	0,1357
	ME	0,6909	0,0942	0,1658
	ME&ICL	0,6885	0,0939	0,1652
20	ICL	0,2030	0,1005	0,1345
	ME	0,6932	0,0917	0,1620
	ME&ICL	0,6916	0,0918	0,1620
30	ICL	0,2026	0,0999	0,1338
	ME	0,6939	0,0906	0,1602
	ME&ICL	0,6913	0,0902	0,1596
50	ICL	0,2026	0,0980	0,1321
	ME	0,6939	0,0891	0,1580
	ME&ICL	0,6912	0,0890	0,1577
70	ICL	0,2028	0,0970	0,1312
	ME	0,6944	0,0888	0,1574
	ME&ICL	0,6914	0,0886	0,1570
100	ICL	0,2027	0,0963	0,1305
	ME	0,6936	0,0881	0,1564
	ME&ICL	0,6907	0,0878	0,1559

Tabla 8: Resultados del método no supervisado utilizando las medidas ICL, ME, la unión de estas ME&ICL y diferentes números de cluster (NC) empleando el algoritmo *GRAPH*

De los resultados de la tablas 5, 6, 7 y 8 observamos un *recall* elevado debido a la sobreestimación que hace el método de lo considerado como plagio. Es decir, hemos declarado más secciones plagiadas de las que realmente hay. Por esto mismo la *precisión* es baja.

Adicional a estas pruebas, se realizaron otros experimentos variando el parámetro del porcentaje mínimo de cubrimiento del documento asignado a la clase NP durante

el proceso de asignación de clases a los diferentes grupos. Los resultados de estos experimentos no han sido reportados explicitamente, debido a que no han habido cambios significativos.

De estos resultados cabe detallar que los resultados obtenidos por la nueva medida propuesta en este trabajo son similares a los resultados obtenidos por las medidas estilísticas, que han sido investigadas desde 1940 y han sido aceptadas por las comunidades de lingüistas como medidas eficaces para medir los cambios de estilo de los autores.

La mayoría de las veces, la unión *ME&ICL* de las dos medidas estílisticas *ICL y ME*, permite obtener buenos resultados sin necesidad de saber a priori si se trata de uno de los casos en donde *ICL* o *ME* funciona mejor. Por el contrario, la unión *ME&ICL* obtiene peores resultados que los obtenidos por las medidas *ICL* o *ME*, cuando el número de cluster *NC* aumenta, esto es debido a que al haber mayor número de cluster agrupar fragmentos con cohesión estilística es más díficil.

Parte V

Conclusiones

Conclusiones y trabajos futuros

5 Conclusiones y trabajos futuros

5.1 Recapitulación

En esta tesina abordamos el problema de la detección intrínseca de plagio mediante dos métodos, un método supervisado que está basado en el uso de medidas estilísticas y un método no supervisado que aborda este problema mediante el uso de algoritmos de clustering.

Las medidas estilísticas capturan las diferencias de estilo de los autores en los fragmentos del documento, permitiendo obtener los fragmentos anómalos que posiblemente fueron plagiados. Dentro de las medidas empleadas, se encuentra una nueva medida llamada Índice general de concordancia local o IGCL, que captura las características de estilo del documento en su globalidad y la particularización de esta medida la llamamos Índice de concordancia local o ICL, que captura las características de los fragmentos del documento en su contexto adyacente.

El análisis de aplicabilidad del método supervisado se ha realizado con el corpus utilizado en la competición PAN. Para realizar los experimentos en el método supervisado, se han aplicado las medidas estilítiscas y el IGCL al conjunto de documentos. A este conjunto se le ha aplicado la desviación estándar y a su vez se ha realizado el promedio de éstas de manera independiente, para ver la eficiencia de discriminar los fragmentos plagiados de los no plagiados.

En el método no supervisado, usamos el clustering junto con las medidas estilísticas para la agrupación de los fragmentos con similitudes estilísticas y así poder discriminar los fragmentos plagiados de los que no lo son. La eficiencia del método no supervisado se ha realizado mediante el evaluador de la competición del PAN. En este método usamos clustering (agrupamiento) basándonos en las medidas estilísticas y el ICL para caracterizar a un documento como un conjunto de atributos y así poder agrupar los fragmentos con similitudes estilísticas y discriminar los fragmentos sospechosos de plagio. Para este análisis se han hecho dos tipos de pruebas: una de las pruebas consiste en realizar el clustering (agrupamiento) con número determinado de cluster, posteriormente se asigna

el cluster con mayor número de fragmentos a la clase de los No Plagiados y a los 35 restantes a la clase Plagiados; la otra prueba es similar pero a la clase No Plagiados se le asigna los cluster que superen la mitad del tamaño del documento y los restantes a la clase Plagiados.

5.2 Conclusiones

La tarea de la detección intrínseca de plagio es un problema aún abierto y que se está investigando en múltiples ámbitos, debido a la necesidad de desarrollar herramientas automatizadas para detectar el plagio. A lo largo de este trabajo, se demostró el nivel de eficiencia de las medidas estilísticas, en la diferenciación del estilo de los fragmentos de un documento para encontrar los posibles fragmentos plagiados. Los resultados de la nueva medida propuesta en esta tesina, el Índice de Concordancia Local son tan buenos como los obtenidos por las medidas estilísticas tradicionales, que han sido investigadas desde hace más de medio siglo y ratificadas por la comunidad de expertos lingüistas.

El ICL mide la cohesión de los fragmentos sospechosos del documento, por lo tanto, es independiente de la longitud de los documentos. Esto proporciona una ventaja importante en las situaciones donde no disponemos de gran cantidad de información para modelar el estilo de los autores, como sucede en la detección intrínseca de plagio.

Tras la experimentación, pudimos observar que las diferentes medidas estilísticas tradicionales tienen diferente eficacia en la detección intrínseca de plagio, debido a la variedad de las características que capturan del documento. Las mejores características para discriminar los fragmentos plagiados del resto del documento, son la legibilidad del documento y las relacionadas con la riqueza del vocabulario en un texto.

En general, se ha observado que los métodos de clustering dividen mejor al documento entre menor sea el número de grupos. De igual manera, los mejores algoritmos son el RB y el DIRECT, los cuales se basan en divisiones binarias sucesivas por tanto rápidamente descarta una gran parte del documento como material libre de plagio, de forma similar a cuando el número de grupos es pequeño.

Finalmente, aunque la métrica estilística propuesta para discriminar los fragmentos de un documento, ICL, no es significativamente mejor que el grupo de medidas estilísticas, sí resulta más independiente de la longitud del texto y debido a su naturaleza de similitud local, permite ser más atractiva en dominios en donde el estilo puede no tener un comportamiento estándar, como en algunas obras poéticas. También cabe resaltar que la otra medida estilística propuesta, IGCL, al utilizarse en el método supervisado con un conjunto de documentos no plagiados del mismo dominio como conjunto de entrenamiento, obtiene mejores resultados que el uso de las otras medidas estilísticas.

5.3 Líneas de investigación abiertas

Debido a que los resultados obtenidos no son óptimos, debemos investigar nuevas formas de abordar el problema de la detección intrínseca de plagio. Algunas de las líneas interesantes de futuras investigaciones serían:

1. Explorar mejor la selección de los clusters declarados como plagiados.
2. Incorporar al modelo de clustering la posición en el documento del fragmento caracterizado para que los algoritmos de clustering tiendan a agrupar el estilo en vecindades cercanas en el documento.
3. Utilizar conjuntos de características restringidos a elementos de estilo (palabras vacías) en el cálculo del ICL.
4. Combinar nuevos métodos desarrollados con los propuestos por nosotros para comprobar la eficiencia de esta combinación.
5. Realizar una plataforma Web donde los usuarios tengan acceso a los métodos que hemos descrito en este trabajo.

Bibliografía

[1] Barrón-Cedeño A. "Detección automática de plagio en texto." En: *Tesis fin de máster. Departamiento de sistemas informáticos y computación. Universidad Politécnica de Valencia,* (2008) (vid. págs. 5, 12).

[2] Barrón-Cedeño A. "On the mono- and cross-Language detection of text re-Use and plagiarism." En: *Tesis de Doctorado. Departamento de sistemas informáticos y computación. Universidad Politécnica de Valencia,* (jun. de 2012) (vid. pág. 5).

[3] Honore A. "Some simple measures of richness of vocabulary." En: *Association for Literary and Linguistic Computing,* 7.2 (1979), págs. 172-177 (vid. pág. 15).

[4] Stein B. y zu Eissen S. M. "Intrinsic Plagiarism analysis with meta learning." En: *In Proceedings of the SIGIR WorkShop on PAN,* (2007), págs. 45-50 (vid. pág. 14).

[5] Grozea C. y Popescu M. "Encoplot - Tuned for high recall (also proposing a new plagiarism detection score)." En: *Notebook for PAN at CLEF 2012,* 60.3 (mar. de 2009), págs. 538-556 (vid. pág. 14).

[6] Holmes D. "The evolution of stylometric in humanities scholarship." En: *Literary and Linguistic Computing,* 13.3 (sep. de 1998), págs. 111-117 (vid. pág. 15).

[7] Zou D., Long W. y Ling Z. "A cluster-based plagiarism detection method." En: *Notebook for PAN at CLEF 2010 Labs and Workshops,* (2010) (vid. pág. 14).

[8] Stamatatos E. "A survey of modern authorship attribution methods." En: *Journal of the american society for information science and technology,* 60.3 (mar. de 2009), págs. 538-556 (vid. pág. 11).

[9] Stamatatos E. "Intrinsic plagiarism detection using character n-gram profiles." En: *PAN'09* 7.2 (2009), págs. 38-46 (vid. págs. 16, 19).

[10] Vallés Balaguer E. "Empresa 2.0: Detección de plagio y análisis de opiniones." En: *Tesis fin de máster. Departamento de sistemas informáticos y computación. Universidad Politécnica de Valencia,* (2010) (vid. pág. 14).

[11] Sánchez-Vega F. "Detección automática de plagio basada en la distinción y fragmentación del texto reutilizado." En: *Tesis fin de máster. Coordinación de Ciencias Computacionales, Instituto Nacional de Astrofisica óptica y Electrónica,* (ene. de 2011). http://ccc.inaoep.mx/ villasen/tesis/TesisMaestria-FernandoSanchez.pdf (vid. pág. 11).

[12] Sánchez-Vega F. "Representación secuencial enriquecida con información para la detección de plagio." En: *Propuesta de doctorado, Instituto Nacional de Astrofisica óptica y Electrónica, México,* (2012) (vid. pág. 3).

[13] Oberreuter G., L'Huillier G., Ríos S. A. y Velásquez J. D. "Approaches for intrinsic and external plagiarism detection." En: *Notebook for PAN at CLEF 2011,* (2011) (vid. pág. 14).

[14] Yule G. *The Statistical Study of Literary Vocabulary.* Cambridge University Press, 1944 (vid. pág. 15).

[15] Parvatti I. y Abhipsita S. "Document similarity analysis for a plagiarism detection system." En: *In Proceedings of the 2nd Indian Int. Conf. on Artificial Intelligence(IICaI-2005),* (2005), págs. 2534-2544 (vid. pág. 4).

[16] Diederich J. "Computational methods to detect plagiarism in assessment." En: *Proceedings of the 7th International Conference on Information Technology Based Higher Education and Training (ITHET '06),* (2006), págs. 147-154 (vid. pág. 14).

[17] Kasprzak J. y Brandejs M. "Improving the reliability of the plagiarism detection system." En: *Notebook for PAN at CLEF 2010 Labs and Workshops,* (2010) (vid. pág. 14).

[18] Seaward L. y Matwin S. "Intrinsic plagiarism detection using complexity analysis." En: *SEPLN 2009 Workshop PAN,* (2009), págs. 56-61 (vid. pág. 16).

[19] Real Academia de la Lengua Española. "Diccionario de la Lengua Española. Vigésima segunda edición." En: (2008) (vid. pág. 3).

[20] Kestemont M., Luyckx K. y Daelemans W. "Intrinsic plagiarism detection using character trigram distance scores." En: *Notebook for PAN at CLEF 2011,* (2011) (vid. pág. 16).

[21] Koppel M., Schler J. y Argamon S. "Computational Methods in Authorship Attribution." En: *Journal of the american society for information science and technology,* 60.1 (ene. de 2009), págs. 9-26 (vid. pág. 11).

[22] Muhr M., Kern R., Zechner M. y Granitzer M. "External and intrinsic plagiarism detection using a cross-lingual retrieval and segmentation system." En: *Notebook Papers for PAN at CLEF 2010 Labs and Workshops,* (2010) (vid. pág. 16).

[23] Potthast M., Barrón-Cedeño A., Stein B. y Rosso P. "An evaluation framework for plagiarism detection." En: *Proc. of the 23rd Int. Conf. on Computational Linguistics, COLING-2010, Beijing, China,* 23-27 (ago. de 2010), págs. 997-1005 (vid. pág. 28).

[24] Clough P. "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) (vid. pág. 12).

[25] Juola P. "Authorship attribution." En: *In foundations and trends in information retrieval,* 1.3 (dic. de 2006) (vid. pág. 11).

[26] Flesch R. "A new readibility yardstick." En: *Journal of Applied Psychology,* 32 (1948), págs. 221-233 (vid. pág. 15).

[27] Gunning R. *The Technique of Clear Writing.* McGraw-Hill, 1952 (vid. pág. 15).

[28] Rao S., Gupta P., Singhal K. y Majumder P. "External and intrinsic plagiarism detection: VSM and discourse markers based approach." En: *Notebook for PAN at CLEF 2011,* (2011) (vid. págs. 14, 16).