



**UNIVERSIDAD DE CHILE**  
**FACULTAD DE CIENCIAS FISICAS Y MATEMATICAS**  
**DEPARTAMENTO DE INGENIERIA INDUSTRIAL**

MODELO DE FUSION DE SCORE UTILIZANDO TEORÍA DE LA INFORMACIÓN  
PARA INTEGRACIÓN DE SISTEMAS DE SIMILITUD DE DOCUMENTOS

TESIS PARA OPTAR AL GRADO DE MAGÍSTER EN GESTION  
DE OPERACIONES

GERARDO MANUEL GUERRERO QUICHIZ

SANTIAGO DE CHILE

2011



**UNIVERSIDAD DE CHILE  
FACULTAD DE CIENCIAS FISICAS Y MATEMATICAS  
DEPARTAMENTO DE INGENIERIA INDUSTRIAL**

**MODELO DE FUSION DE SCORE UTILIZANDO TEORÍA DE LA INFORMACIÓN  
PARA INTEGRACIÓN DE SISTEMAS DE SIMILITUD DE DOCUMENTOS**

**TESIS PARA OPTAR AL GRADO DE MAGÍSTER EN GESTION  
DE OPERACIONES**

**GERARDO MANUEL GUERRERO QUICHIZ**

**PROFESOR GUIA:  
SEBASTIAN RIOS PEREZ**

**MIEMBROS DE LA COMISION:  
JUAN VELASQUEZ SILVA  
GASTON L'HUILLIER CHAPARRO  
JUAN MARIN CAIHUAN**

**SANTIAGO DE CHILE  
JUNIO, 2011**

*“If we all did the things we are capable of doing,  
we would literally astound ourselves.”*

*–Thomas A. Edison  
US inventor (1847 - 1931)*

# Resumen Ejecutivo

El presente proyecto de tesis se enmarca dentro del proyecto *FONDEF DO8I-1015 llamado Document Copy Detector (DOCODE)*, cuyo objetivo es desarrollar un sistema de detección de copia escrita. Hoy ya existe una versión 1.0 de DOCODE que incluye búsqueda web a partir de un texto ingresado<sup>1</sup>, sin embargo se está desarrollando la versión 2.0 en la cual se incluirán parseadores, *sistemas de similitud de documento*, y demás herramientas avanzadas; y es en esta nueva versión que se incluirá el *Modelo de Fusión de Datos* que aquí se describe.

Antes de explicar el desarrollo de esta tesis recordemos la etapa de cambio que vivimos en la actualidad, la llamada *revolución informática*, que en pocas palabras es la masificación y facilidad de acceso a la información mediante equipos electrónicos. Esta etapa de cambio se ve fortalecida con el Internet, medio que permite a las personas consultar e intercambiar información con terceros con bastante facilidad. Esta comodidad de acceso a la información también trae consigo un inconveniente: el problema del plagio, un mal que en esta época de adelantos puede traer atrasos, porque un estudiante en formación en lugar de investigar se puede dedicar a copiar y pegar información que encuentra fácilmente en Internet. Este problema no es menor y repercute en muchos ámbitos, no sólo académico, porque además de ser un problema ético, a gran escala se puede convertir en un problema serio con índices legales.

Para evitar ello, los investigadores del tema han desarrollado diversos métodos y sistemas de detección de plagio. Los cuales se basan en metodologías o algoritmos numérico-matemáticos que ayudan a identificar el grado de similitud entre un par de documentos A y B, también denominado dupla de *Documento Sospechoso vs. Documento Fuente*. Estos desarrollos poseen un variado desempeño, y es dependiente de la base de prueba. Es decir, algunos métodos funcionan bien y dan un resultado confiable para cierta base de experimentación, pero otros no, y estos mismos métodos pueden entregar malos resultados para otra base, mientras que los otros dan buenos resultados. Eliminar la “*incertidumbre*” en los resultados es la motivación principal de esta tesis, por ello se propone desarrollar un modelo para detección de plagio que pueda incluir  $N$  métodos de detección individuales (Donde:  $N \in \mathbb{Z}^+$ ) y que sea capaz de tomar sus mejores resultados para mostrarlos como un único resultado final.

Con lo descrito, se plantea el objetivo de este proyecto: “Desarrollar un *Modelo de Fusión de Datos* eficiente que pueda integrar diversos resultados de *Sistemas de detección de similitud entre documentos*”

Para conseguir dicho objetivo se *Diseñó y Desarrolló un Modelo de Fusión de Datos* para la detección de plagio entre documentos que posee tres partes importantes: (1) La modificación de la *Ecuación del Valor de la Información* propuesta por Yu Suzuki et. al. [67]. (2) Un *Sistema de Combinación Geométrico* y (3) Una formulación que incluye un *Factor de credibilidad*. Que es un indicador ingresado por el usuario (juicio experto) y que muestra el nivel de confianza que se le tiene a un Método de Detección de Plagio.

Posteriormente, el *Modelo propuesto* se validó con una base de pruebas supervisada otorgada por la PAN2010<sup>2</sup> [55] y se le comparó con otros *Modelos de Fusión de Datos Clásicos* [49, 63]. En esta comparación el *Modelo de Fusión de Datos Propuesto en la tesis* alcanzó el mejor desempeño con un F-MEASURE promedio de 94.3 % y una desviación estándar de 8.2 %, logrando así ser el más eficiente entre los modelos.

Además, con ayuda del grupo de Social Network Analysis (SNA) de la Universidad de Chile<sup>3</sup>, se realizó un análisis para detectar grupos sociales de copia para un conjunto de tareas digitales presentadas por alumnos del ramo de Tecnologías de la Información<sup>4</sup> donde se logró detectar relación de similitud entre tareas de algunos alumnos. Esto después se contrastó con el auxiliar del ramo y se verificó la existencia de copia para los documentos reconocidos por el sistema. En esta etapa se utilizaron grafos dirigidos, para la representación visual de los resultados.

Finalmente, se concluyó que el sistema desarrollado es eficiente con un ACCURACY, PRECISION y RECALL de 99.8 %, 96.1 % y 78.1 % respectivamente. Consiguendo, de ese modo, cumplir con el objetivo propuesto.

<sup>1</sup>www.docode.cl. Último acceso: 21-Ene-2011

<sup>2</sup>Base de entrenamiento supervisada, libre y publicada por el Workshop internacional de Detección de Plagio Escrito PAN2010

<sup>3</sup>Grupo creado a partir del proyecto FONDECYT de iniciación (ID: 11090188)

<sup>4</sup>Código IN72K. Dictado en el 2010 por el PhD. Sebastián Ríos P.

# Agradecimientos

Este documento representa el fin de una etapa que comenzó por allá en Enero del 2008, cuando gracias a CONICYT<sup>5</sup> fui seleccionado como becario para una maestría en Chile. Entre alegría y desconfianza por lo que involucraba salir de Perú a un entorno nuevo, lleno de expectativas por el estudio y la oportunidad que tenía. No me imaginaba todo lo que sería esta grata experiencia en Chile. Primero el conocer a Gabriela Cano, quien además de ser mi novia es mi gran amiga y compañera. Luego, personas muy valiosas como son el Sr. Pedro Cifuentes, la Sra. Elida Velásquez y su hija Natalia Tapia quienes fueron las primeras personas aquí en Chile, que prácticamente sin conocerme me abrieron las puertas de su casa y con gran hospitalidad hicieron que sienta en Santiago mi segundo hogar.

En el entorno de la universidad, fue un gusto compartir aulas con destacados alumnos. Y, a diferencia de lo que traía como pre-juicio, siempre fueron bastante cercanos y buena onda; haciéndome sentir como parte del grupo. Lo agradable que se sentía el recibir cátedras de grandes profesionales, expertos en sus áreas. Hasta el nivel académico exigido por la Universidad que me hizo dudar más de una vez si yo sería capaz de continuar con el Magíster, y que también me hizo sentir el dolor de fallar a pesar de esforzarse tanto –creo que todos estos factores son los que le agregaron el valor que hoy en día encuentro a los estudios–.

No imaginaba que, a casi dos años y medio de mi llegada a Chile, me pudiese sentir tan a gusto y feliz de haber venido aquí a pasar esta etapa de mi vida –o del inicio de mi vida como diría mi padre–. De haber tenido la oportunidad de aprender, de haber conocido el amor, la amistad y también de demostrar mis habilidades como profesional. Este proceso del Magíster tal vez se cierra ahora con esta tesis, pero otras nuevas etapas se abren y con ellas vienen nuevos retos para los que ahora me siento preparado.

Quisiera empezar mis agradecimientos dando una gran mención a mis padres Pedro Guerrero y Beatriz Quichiz porque gracias a ellos he conocido lo que es tener coraje para afrontar todos los retos de la vida. Ambos, desde mi etapa de pre-grado dieron todo de su parte para que yo consiguiera mis estudios. En verdad, ¡muchas gracias mis viejitos! yo siempre recordaré y tendré presente esos esfuerzos que hacían, que hasta se quedaban sin comer para que yo siguiera mis estudios. Por darme todo a mi se enfermaron, yo no tengo palabras, ni medios para agradecerles por todo –y hasta al redactar esto me salen algunas lágrimas– pero les prometo que nunca les voy a fallar y que a partir de ahora yo velaré por ustedes para que todo este siempre bien en casa. Papá y mamá, estoy muy orgulloso de ustedes, tal vez la frase queda corta, pero *Gracias por todo*.

Mis familiares cercanos: mi tía Lidia, Sulme, Gustavo, Glenn y mis primos Julio, Carlos, Gustavo, Sergio, Sandra, Juan Carlos y Luis. Que a la distancia me apoyaban con buenos deseos, y que a cada visita a Lima siempre estaban presentes para reunirnos y compartir momentos en familia. Gracias por acompañar y visitar a mis padres cuando los dejé solos en casa. Especialmente gracias por desordenar un poco la casa, mi mamá se alegraba de acordarse de poner las cosas en su lugar, porque le hacía pensar que yo estaba en casa.

Alguien que no alcanzó a ver este documento completo, y a quien no pude acompañar en su despedida, es a mi tío Manuel Quichiz Q.E.P.D. (9 de Junio del 2009). Él también era un familiar cercano pero por

---

<sup>5</sup>Comisión Nacional de Investigación Científica y Tecnológica de Chile

---

cosas de la vida, ya no está con nosotros. Tío, no logramos coincidir en tiempos y soy consciente que no me podrá leer pero de todas maneras quiero agradecerle y disculparme de no haber estado ahí en su despedida, este documento representa el motivo por el cual no pude estar ahí, yo sé que usted hubiese entendido.

También quiero dar un especial agradecimiento a mi novia por haber compartido su tiempo conmigo, por haber soportado en ocasiones mi mal genio y por haberme apoyado en todo este proceso. Gracias *pinolillerita*, por aportar con tus locuras alegría en mis días, por ayudarme a ordenarme, por cocinar rico y –aunque yo reniegue– por controlarme cuando quiero malgastar dinero y hasta por ser celosa, porque hasta enojada te ves hermosa. Eres una persona excepcional, me encantará seguir compartiendo nuestras vidas en el futuro, te amo.

Llegando a los agradecimientos en Chile, como ya lo adelantaba al Sr. Pedro y a la Sra. Ely muchas gracias por ser como mis segundos padres. O mejor dicho, mis padres aquí en Chile, por preocuparse por mí, por se tan cariñosos conmigo. Siempre tendrán un lugar especial en mi vida.

En el contexto académico, agradecer a mis profesores y amigos Sebastián Ríos y Juan Velásquez que han sido mis guías en las investigaciones de tesis, gracias por la confianza depositada en mí. Por las oportunidades que me ofrecieron y porque me hicieron sentir como uno más del equipo Chileno de desarrollo. Gracias profes, muchos de los logros profesionales conseguidos en Chile fueron gracias a las oportunidades que me brindaron. Nunca aprendí muchas palabras en japonés pero, y espero no equivocarme al escribir, “*Arigatou*”.

A los muchachos del DOCODE: Gastón L’Huillier quien ha sido un guía y amigo, también a los demás *cabros* –expresión que se usa en Chile para referir “amigo”–: Héctor, Felipe, Gabriel, Eduardo, Rodrigo y particularmente a Patricio (el Pato Moya) con quienes compartí buenos momentos mientras trabajábamos para el proyecto. Aún nos debemos un asado y algunas visitas a comer comida peruana.

No quiero dejar de agradecer al proyecto FONDEF Código: DO8I-1015, titulado DOCument COpy DETector (DOCODE) por permitirme ser parte de su equipo de investigadores. Además, mi cordial gratitud al Instituto Sistemas Complejos de Ingeniería CM: P-05-004-F y CONICYT: FB016 centro que nos apoyó con un espacio para investigar en la Universidad de Chile.

Ya para finalizar, retomando las palabras que utilicé cuando iniciaba los agradecimientos, esta tesis cierra este grato proceso que fue llegar a Chile e iniciar el Magister en Gestión de Operaciones. Un proceso que me ha enseñado a valorar mucho a las personas, que en cierto modo me ayudó a madurar, que me ha enseñado sobre la fortaleza que había que tener, que me enseñó de mí mismo y sobre todo que me ha dado la visión necesaria para aspirar a más, para ponerme retos mayores y saber que tengo la capacidad de ir por ellos. El viaje a Ítaca valió la pena.

Mencioné esta última frase porque a pocas semanas de mi llegada a Chile, por cuestiones del azar un texto llegó a mis manos y como si fuese algún mensaje dirigido, entre metáforas, mencionaba todo aquello que debía esperar de este viaje a Chile –mi viaje a Ítaca–.

---

## ÍTACA (1911)

Cuando emprendas tu viaje a Ítaca  
pide que el camino sea largo,  
lleno de aventuras, lleno de experiencias.  
No temas a los Lestrigones ni a los Cíclopes,  
ni al colérico Poseidón,  
seres tales jamás hallarás en tu camino,  
si tu pensar es elevado, si selecta  
es la emoción que toca tu espíritu y tu cuerpo.  
Ni a los Lestrigones ni a los Cíclopes  
ni al salvaje Poseidón encontrarás,  
si no lo llevas dentro de tu alma,  
si no los yergue tu alma ante ti.

Pide que el camino sea largo.  
Que sean muchas las mañanas de verano  
en que llegues -¡con qué placer y alegría!-  
a puertos antes nunca vistos.  
Detente en los emporios de Fenicia  
y hazte con hermosas mercancías,  
nácar y coral, ámbar y ébano  
y toda suerte de perfumes voluptuosos,  
cuantos más abundantes perfumes voluptuosos puedas.  
Ve a muchas ciudades egipcias  
a aprender de sus sabios.

Ten siempre a Ítaca en tu pensamiento.  
Tu llegada allí es tu destino.  
Mas no apresures nunca el viaje.  
mejor que dure muchos años  
y atracar, viejo ya, en la isla,  
enriquecido de cuanto ganaste en el camino  
sin aguardar a que Ítaca te enriquezca.

Ítaca te brindó tan hermoso viaje.  
Sin ella no habrías emprendido el camino.  
Pero no tiene ya nada que darte.

Aunque la halles pobre, Ítaca no te ha engañado.  
Así, sabio como te has vuelto, con tanta experiencia,  
entenderás ya qué significan las Ítacas.

*Konstantinos P. Kavafis*

(29 de Abril, 1893 – 29 de Abril, 1933)

Traducción de Pedro Bádenas de la Peña. [4]

# Índice general

<b>Resumen</b>	<b>I</b>
<b>Agradecimientos</b>	<b>II</b>
<b>Índice general</b>	<b>V</b>
<b>Índice de cuadros</b>	<b>IX</b>
<b>Índice de figuras</b>	<b>XI</b>
<b>1 INTRODUCCIÓN</b>	<b>1</b>
1.1 Antecedentes generales . . . . .	1
1.2 Objetivos . . . . .	4
1.2.1 Objetivo general . . . . .	5
1.2.2 Objetivos específicos . . . . .	5
1.3 Resultados Esperados . . . . .	5
1.4 Metodología . . . . .	6
1.5 Estructura de la tesis . . . . .	7
<b>2 TRABAJO PREVIO</b>	<b>8</b>
2.1 El problema del plagio . . . . .	8
2.2 Tipos de plagio . . . . .	9
2.2.1 Copia por intención . . . . .	10
2.2.2 Copia por estructura . . . . .	12
2.2.3 Copia por el origen de la fuente . . . . .	13



2.2.4	Colusión . . . . .	14
2.3	Métodos para detección de plagio . . . . .	15
2.3.1	Métodos manuales de detección de plagio . . . . .	16
2.3.2	Métodos numéricos para detección de plagio . . . . .	17
2.3.3	Software para detección de plagio . . . . .	18
2.4	Métodos de fusión de datos . . . . .	22
2.4.1	Beneficios de un método de fusión de datos . . . . .	25
2.4.2	Métodos de fusión de datos basados en ranking . . . . .	27
2.4.3	Métodos de fusión de datos basados en score . . . . .	36
<b>3</b>	<b>MODELO DE FUSIÓN DE DATOS POR SCORE</b>	<b>40</b>
3.1	Notaciones para el capítulo . . . . .	41
3.2	Consideraciones previas . . . . .	41
3.3	Ecuación del Valor de la Información Modificada . . . . .	45
3.4	Sistema de Combinación Geométrica . . . . .	47
3.5	Factor de Credibilidad . . . . .	49
3.6	Modelo de Fusión de Datos por Score . . . . .	50
3.7	Algoritmo del Modelo de Fusión de Datos por Score . . . . .	54
<b>4</b>	<b>EXPERIMENTACIONES</b>	<b>61</b>
4.1	Selección de datos . . . . .	62
4.1.1	PAN plagiarism corpus . . . . .	63
4.1.2	Lógica de selección de datos . . . . .	63
4.2	Cálculo de Scores . . . . .	65
4.2.1	SimParalelo . . . . .	65
4.2.2	SimTFIDF . . . . .	66
4.2.3	SimAlfabético . . . . .	68
4.2.4	Diff . . . . .	68
4.3	Modelo de Fusión de Datos . . . . .	70
4.4	Punto de Corte . . . . .	71
4.5	Comparación de resultados . . . . .	71

<b>5 ANÁLISIS DE RESULTADOS</b>	<b>76</b>
5.1 Resultados del Punto de Corte óptimo . . . . .	76
5.2 Resultados del p-segmento óptimo . . . . .	78
5.3 Reconstrucción del número de Fuentes para cada Documento Sospechoso . . . . .	79
5.4 Métodos de Detección de Similitud . . . . .	81
5.5 Comparación con otros Modelos de Fusión de Datos . . . . .	83
5.6 Análisis de Colusión . . . . .	88
<b>6 CONCLUSIONES Y TRABAJO FUTURO</b>	<b>91</b>
6.1 Conclusiones de la tesis . . . . .	91
6.2 Trabajo futuro . . . . .	93
6.2.1 Linealización . . . . .	93
6.2.2 Cálculo de la Frecuencia . . . . .	94
6.2.3 Cálculo del Valor de la Información modificada . . . . .	94
<b>REFERENCIAS</b>	<b>96</b>
<b>Apéndice A CONCEPTOS DE TEORÍA DE LA INFORMACIÓN</b>	<b>103</b>
A.1 Información . . . . .	103
A.2 Teoría de la Información . . . . .	104
A.3 Entropía . . . . .	105
<b>Apéndice B INDICADORES DE EFICIENCIA EN RECUPERACIÓN DE LA INFORMACIÓN</b>	<b>106</b>
B.1 Precision . . . . .	107
B.2 Recall . . . . .	107
B.3 Accuracy . . . . .	107
B.4 F-measure . . . . .	108
<b>Apéndice C ANÁLISIS DE LA LINEALIZACIÓN DE SCORES – YU SUZUKI ET. AL.</b>	<b>109</b>
<b>Apéndice D TABLAS DE VALORES RESULTANTES</b>	<b>113</b>
D.1 Lista de Documentos seleccionados del PAN Corpus 2010 . . . . .	113

D.2	Análisis para elección del punto de corte óptimo . . . . .	114
D.2.1	Corte entre 0 y 1 . . . . .	114
D.2.2	Corte entre 0 y 0.004 . . . . .	115
D.2.3	Corte entre 0.00022 y 0.00042 . . . . .	115
D.2.4	Corte entre 0.0003122 y 0.0003922 . . . . .	115
D.3	Análisis para elección del p-segmento óptimo . . . . .	116
D.3.1	p-segmento entre 2 y 1000 . . . . .	116
D.3.2	p-segmento entre 400 y 600 . . . . .	116
D.4	Indicadores para todos los documentos seleccionados . . . . .	117

# Índice de cuadros

Cuadro 2.1	Cotas para el cuantificador difuso $Q$ . . . . .	33
Cuadro 3.1	Matriz genérica de Scores . . . . .	43
Cuadro 3.2	Vector de Scores Fusionados . . . . .	43
Cuadro 3.3	Vector de Frecuencias de Scores . . . . .	56
Cuadro 4.1	Documentos Sospechosos y el Número de Documentos Fuentes que contienen . . . . .	65
Cuadro 4.2	Variaciones del SimParalell . . . . .	67
Cuadro 4.3	Variaciones del SimTFIDF . . . . .	67
Cuadro 4.4	Variaciones del SimAlfabético . . . . .	68
Cuadro 4.5	Variaciones del Diff . . . . .	69
Cuadro 4.6	Factor de Credibilidad para los Método de Similitud de Documentos . . . . .	70
Cuadro 4.7	Modelos clásicos de fusión por Score . . . . .	72
Cuadro 4.8	Valores del factor $t$ . . . . .	75
Cuadro 5.1	Indicadores para los valores óptimos . . . . .	79
Cuadro 5.2	Indicadores para los valores óptimos . . . . .	81
Cuadro 5.3	Resultados de los Métodos de Similitud en la BD de prueba. . . . .	81
Cuadro 5.4	Modelos clásicos de fusión por Score . . . . .	83
Cuadro 5.5	Indicadores para los Modelos de Fusión Clásicos y para el MetaScore (Modelo propuesto) . . . . .	84
Cuadro 6.1	Valores que puede tomar el ranking $R_k(x_i)$ . . . . .	94
Cuadro 6.2	Valor de la Información para los modelos de Yu Suzuki et. al. [67], propuesto en la tesis y el Trabajo Futuro . . . . .	95

Cuadro B.1 Matriz de confusión . . . . .	106
Cuadro C.1 Test 1 de la técnica de linealización Yu Suzuki et. al. . . . .	110
Cuadro C.2 Test 2 de la técnica de linealización Yu Suzuki et. al. . . . .	111
Cuadro D.1 Lista de los 80 Documentos seleccionados desde el PAN Corpus 2010 . . . . .	114
Cuadro D.2 Lista de Documentos Sospechosos y sus respectivos Documentos Fuentes . . . . .	121
Cuadro D.3 Rango para el corte óptimo entre 0 y 1 . . . . .	121
Cuadro D.4 Resultados para el corte óptimo entre 0 y 1 . . . . .	122
Cuadro D.5 Rango para el corte óptimo entre 0 y 0.004 . . . . .	123
Cuadro D.6 Resultados para el corte óptimo entre 0 y 0.004 . . . . .	124
Cuadro D.7 Rango para el corte óptimo entre 0.00022 y 0.00042 . . . . .	125
Cuadro D.8 Resultados para el corte óptimo entre 0.00022 y 0.00042 . . . . .	126
Cuadro D.9 Rango para el corte óptimo entre 0.0003122 y 0.0003922 . . . . .	127
Cuadro D.10 Resultados para el corte óptimo entre 0.0003122 y 0.0003922 . . . . .	128
Cuadro D.11 Rango para el p-segmento óptimo entre 2 y 1000 . . . . .	129
Cuadro D.12 Resultados para el p-segmento óptimo entre 2 y 1000 . . . . .	129
Cuadro D.13 Rango para el p-segmento óptimo entre 400 y 600 . . . . .	130
Cuadro D.14 Resultados para el p-segmento óptimo entre 400 y 600 . . . . .	130
Cuadro D.15 Indicadores para el $corte = 3,5 \cdot 10^{-4}$ y $p-segmento = 500$ . . . . .	132

# Índice de figuras

Figura 1.1	Técnica de similitud entre documentos . . . . .	2
Figura 1.2	Sistema de Fusión de Datos . . . . .	3
Figura 2.1	Taxonomía del plagio [26] . . . . .	12
Figura 2.2	Red Social de Copia - 2 individuos . . . . .	14
Figura 2.3	Red Social de Copia - 13 individuos . . . . .	15
Figura 2.4	Sistema de similitud de documentos (doc. sospechoso vs. docs. fuentes . . . . .	22
Figura 2.5	Relación entre los métodos de similitud de <i>Score</i> y de <i>Ranking</i> . . . . .	23
Figura 2.6	Sistemas de fusión de datos . . . . .	24
Figura 2.7	Sistemas de detección . . . . .	25
Figura 2.8	Función para el cuantificador difuso Q . . . . .	33
Figura 2.9	Histograma de la linealización de Yu Suzuki et.al. [67] . . . . .	38
Figura 3.1	Matriz de Scores - Detalle por página . . . . .	42
Figura 3.2	Método de Detección de Similitud . . . . .	42
Figura 3.3	Matriz de Scores - Por documento completo . . . . .	43
Figura 3.4	Sistemas de fusión de datos . . . . .	50
Figura 3.5	Frecuencia que los Scores se encuentren en un Segmento r-ésimo . . . . .	51
Figura 4.1	Proceso de Detección de Plagio de Documentos (Proceso DPD) . . . . .	61
Figura 4.2	Distribución de los documentos del PAN plagiarism corpus 2010 . . . . .	63
Figura 4.3	Histograma de los Documentos Fuentes y Sospechosos del PAN corpus 2010 . . . . .	64
Figura 4.4	Histograma de los Documentos Fuentes y Sospechosos seleccionados . . . . .	64
Figura 4.5	Detalle del funcionamiento del Método de Similitud . . . . .	66

Figura 5.1	Media de los indicadores para el Punto de Corte óptimo . . . . .	77
Figura 5.2	Desviación estándar de los indicadores para el Punto de Corte óptimo . . . . .	77
Figura 5.3	Media de los indicadores para el P-segmento óptimo . . . . .	78
Figura 5.4	Desviación estándar de los indicadores para el P-segmento óptimo . . . . .	78
Figura 5.5	Reconstrucción del histograma del PAN Corpus 2010 . . . . .	79
Figura 5.6	Comparación de histograma supervisado y predicho por el Método de Fusión de Scores	80
Figura 5.7	Modelo de Fusión de Datos con valores filtrados . . . . .	80
Figura 5.8	Documentos Fuentes vs. Documento Sospechoso por Método de similitud y de Fusión (escala logaritmica) . . . . .	82
Figura 5.9	Documentos Fuentes vs. Documento Sospechoso por Método de similitud y de Fusión	82
Figura 5.10	Comparación de Modelos de Fusión . . . . .	83
Figura 5.11	Comparación de Modelos por el ACCURACY . . . . .	85
Figura 5.12	Comparación de Modelos por la PRECISION . . . . .	85
Figura 5.13	Comparación de Modelos por el RECALL . . . . .	86
Figura 5.14	Comparación de Modelos por el F-MEASURE . . . . .	86
Figura 5.15	Lógica de análisis de copia entre Documentos para las tareas del IN72K . . . . .	88
Figura 5.16	Análisis de Colusión para la tarea 1 del IN72K . . . . .	89
Figura 5.17	Análisis de Colusión para la tarea 2 del IN72K . . . . .	90
Figura 6.1	Curvas del Valor de la Información para los modelos de Yu Suzuki et. al [67], pro- puesto en la tesis y el Trabajo Futuro . . . . .	95
Figura A.1	Modelo del sistema de comunicación . . . . .	104
Figura B.1	Diferencia entre Accuracy y Precision . . . . .	108
Figura C.1	Linealización – Técnica Suzuki et. al . . . . .	111
Figura C.2	Linealización – Técnica Suzuki et. al . . . . .	112

# Capítulo 1

## Introducción

En este capítulo, se presentará una breve descripción del problema del plagio y la solución informática a esta problemática. Posteriormente se repasarán antecedentes generales sobre modelos de fusión de datos cuyo objetivo es integrar dos o más algoritmos de *Detección de Similitud entre documentos*. Se discutirá el objetivo general y los objetivos específicos que defiende esta tesis, luego se presentará la metodología utilizada para el desarrollo de la investigación. Finalmente se abordará la estructura de este documento de tesis, dando una breve descripción por capítulo.

### 1.1. Antecedentes generales

Todo sistema para *detección de plagio* busca reconocer si un *Documento Sospechoso* es efectivamente un plagio y además indicar cual es el *Documento Fuente* de donde procede. Para ello se encarga de analizar las duplas *Documento Sospechoso – Documento Fuente* con diversos procedimientos que le ayuden a detectar el nivel de similitud entre documentos, el cual posteriormente ayudará a indicar si hubo plagio, o no. Sin embargo, como ningún sistema es perfecto, muchos de los *Detectores de Similitud* sólo son eficientes para una porción del universo de casos. Es decir, si se tiene un único *Documento Sospechoso* y se le compara contra un conjunto de  $m$  *Documentos Fuentes* muchos de los *Detectores de Similitud* no reconocerán eficientemente todas las duplas *Sospechoso – Fuente* que existan. Con la necesidad de contar con un *Detector de Plagio* que pueda cubrir un mejor rango de detección es que se enmarca este *Proyecto de Tesis* el cual, además, busca minimizar los casos de error que presentan los *Sistemas de Detección de similitud* integrando en un sólo modelo, los resultados de diversos *Algoritmos de Detección*.

Luego del breve repaso acerca de los *Sistemas de Detección de Plagio* y del objetivo que guía la tesis se harán algunas definiciones que han servido de motivación para el desarrollo del proyecto: El *plagio* es la “acción de tomar el trabajo de otra persona como el trabajo de uno mismo”. En el contexto de esta tesis, se hará referencia al plagio como el *uso de los datos de otra persona, entidad o medio sin especificar el*



*origen de su fuente* y se considerará exclusivamente plagio escrito<sup>1</sup>. Además, se entiende que la *copia* es la imitación de una obra ajena, con la pretensión de que parezca original. Debido a la connotación cercana entre definiciones de *copia* y *plagio*, en este documento se utilizarán indistintamente ambas definiciones.

A nivel ético-académico, y sin entrar en la connotación legal que el plagio pueda ocasionar, la copia en cualquiera de sus formas es una falta repudiable y sancionable. Con el fin de resolver el problema del plagio diversas entidades académicas como la Universidad de Oxford [12] y la Universidad del Mississippi del Sur [36] han desarrollado manuales especializados para afrontar el tema; sin embargo, a pesar de sus esfuerzos, y por muy severas que sean las sanciones, el plagio ha seguido presente convirtiéndose en un problema bastante común y conocido por muchos. Sólo en Chile, en un estudio de investigación realizado para el proyecto DOCODE [13, 14], se encontró que alrededor del 79.2 % de alumnos encuestados<sup>2</sup> aceptaron haber incurrido en el delito de plagio al menos una vez en su vida; además afirmaron, en el 93.5 % de las ocasiones, que sus compañeros han copiado al menos una vez en su vida. Al margen de lo peculiar de las respuestas<sup>3</sup>, el estudio nos revela que al menos el 80 % de estudiantes en Chile han plagiado alguna vez en su vida. Estudios, realizados en Estados Unidos [17, 29] y el Reino Unido [75] muestran comportamientos y porcentajes de copia similares.

Ante esta problemática, investigadores del área de las Tecnologías de la Información<sup>4</sup>, particularmente el área de Recuperación de la Información (IR<sup>5</sup>), han abordado el tema y con ello han surgido diversas técnicas para encontrar similitud entre documentos. La lógica de estas técnicas o métodos de similitud (Ver Figura 1.1) consideran como *Entrada (Input)* a dos documentos: *Documento A* y *Documento B*, ambos son descompuestos en partes<sup>6</sup> según el criterio del *Sistema de Detección de Similitud*, luego se evalúan. Finalmente, el *Sistema* entrega un puntaje de *Salida (Output)* que indica la similitud entre el par *Documento A – Documento B*. Para una mejor definición, el *Documento A* se puede considerar como el *Documento Sospechoso* de copia y el *Documento B* como (posible) *Documento Fuente*.

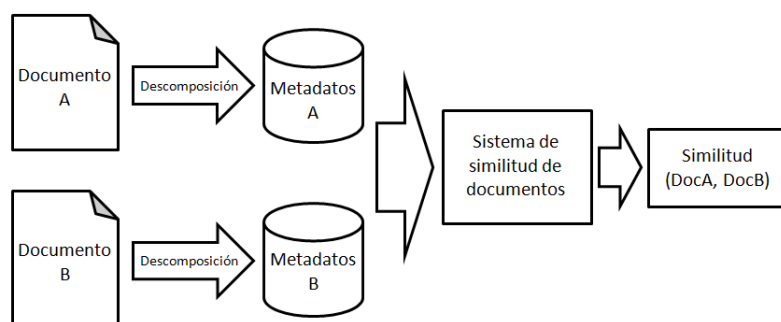


Figura 1.1: Técnica de similitud entre documentos

<sup>1</sup>En el marco legal existe otra variedad de copia, denominada “plagio intelectual” o *robo de ideas*

<sup>2</sup>Se encuestaron a 3200 alumnos y 300 profesores en Chile. El estudio cubrió instituciones de educación como institutos profesionales, universidades y colegios de enseñanza media pertenecientes a las regiones V, VI, VII y Metropolitana.

<sup>3</sup>Como el mismo estudio realizado para DOCODE [14] lo indica, esto se puede deber al concepto Chileno del “*acusador*” donde los alumnos indican un: “*Yo no fui, el otro si*”

<sup>4</sup>También denominado Information Technology, de su nombre en inglés

<sup>5</sup>Information Retrieval, de su nombre en inglés

<sup>6</sup>También definidos *Metadatos*

En la literatura se pueden encontrar desde *Sistemas Básicos* como metodologías para la detección de *Similitud de Documentos* hasta *Sistemas Avanzados* como algoritmos matemáticos que buscan similitudes directas basadas en palabras [58] y n-gramas [18] o agrupaciones de palabras con el Análisis Semántico Latente (LSA<sup>7</sup>) [21, 37]. Sin embargo, cada uno de estos métodos presenta ser efectivo pero para cierto nivel de casos prácticos y no se les puede generalizar porque en general la capacidad humana para emitir juicios de similitud de documentos aún es muy superior. Esto lo confirma Michael Lee en su análisis de métodos de similitud de documentos [40], donde indica que la principal debilidad de los modelos matemáticos es que no han sido contrastados contra la suficiente data empírica para su formulación, por ello su eficacia se ve reducida en casos prácticos.

Para resolver la falta de generalidad surgen modelos que se encargan de **fusionar** los resultados de más de un método de *Similitud de Documentos*. Usualmente los *Modelos de Fusión de Datos* han sido orientados y aplicados a la Web como los denominados *MetaSearches* (o MetaBuscadores, en español) [8, 24, 25, 38, 42, 49, 51, 61, 63, 67, 79] que utilizan los resultados de diversos motores de búsqueda Web y los fusionan para así entregar un único resultado representativo.

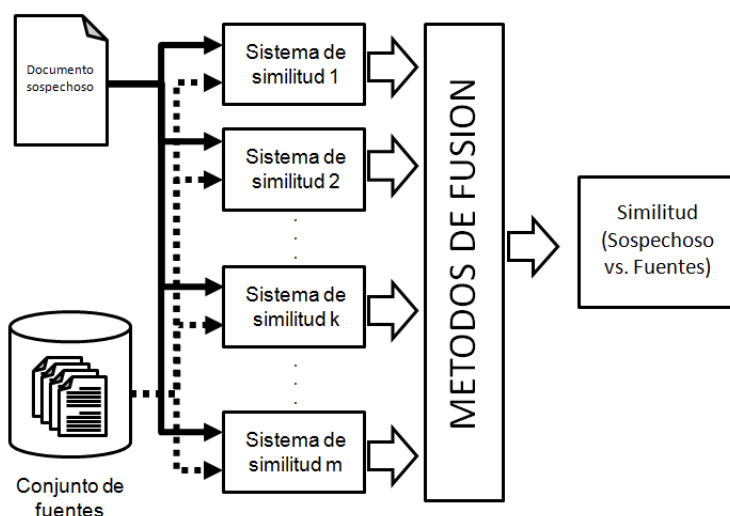


Figura 1.2: Sistema de Fusión de Datos

En el contexto de la detección de plagio, los métodos de fusión de datos son una buena alternativa para generalizar y encontrar una mejor similitud entre documentos (sospechoso y fuentes). Como lo indica Shengli Wu et al. en su trabajo "*Performance prediction of data fusion for information retrieval*"[77], y como también se afirma en estudios previos realizados por Amitay et. al.[1], Soboroff et. al. [66] y Wu et. al. [76, 78].

<sup>7</sup>LSA: Latent Semantic Analysis por su nombre en inglés

Para que un *Método de Fusión de Datos* sea eficiente hay factores que deben ser considerados:

- **Sistemas de Similitud de Documentos con Input similar.**

Al observar la Figura 1.2 se aprecia todos los *Sistemas de Similitud de Documentos* toman como entrada la misma dupla *Documento Sospechoso – Documento Fuente*; como esta entrada es común, y única, debe quedar claramente definido que todos los métodos aceptan dichos datos de entrada. Para el caso particular de la tesis, es indicar que *todos* los documentos acepten como entrada (input) documentos de texto plano.

- **Tiempo computacional de los algoritmos integrados al método de fusión.**

Este factor será de crucial importancia, porque el método de fusión depende directamente de los resultados de las técnicas de similitud de documento e indirectamente de su tiempo computacional. Es decir, el método de fusión de datos no podrá iniciar su proceso hasta que todas las técnicas de similitud de documento hayan culminado. Aquí se podrá elegir entre técnicas que consuman *más tiempo computacional* pero que a la vez entreguen una *buena calidad de resultados* o elegir técnicas que consuman *poco tiempo computacional* pero que otorguen una *baja calidad de resultados*.

- **Método de Fusión de Datos que reciba valores estandarizados.**

Nuevamente, haciendo referencia a la Figura 1.2 se observa que luego del análisis por el *Sistema de Similitud entre Documentos* los resultados van a la entrada del *Método de Fusión de Datos*; este paso de un proceso a otro debe ser regulado, porque nada asegura que todos los *Sistemas de Similitud* entreguen un Output en el mismo rango. Es decir, para algún Sistema la Similitud entre documentos puede ser binaria, entregar un cero cuando no hay similitud y un uno cuando sí la hay; o puede ser por etiquetas de “baja”, “media” y “alta” o un puntaje entre  $- \text{inf}$  y  $+ \text{inf}$ , entre cualquier otra probable variedad. Por ello, este factor se debe considerar para que exista una *etapa intermedia* de transformación y/o estandarización de resultados entre la salida de los *Sistemas de Similitud* y la entrada del *Método de Fusión de Datos*

## 1.2. Objetivos

Como motivación de la tesis, se plantea desarrollar un *Modelo de Fusión de Datos* que permita determinar si un *Documento Sospechoso* tiene alta o baja probabilidad de plagio respecto a un conjunto de *Documentos Fuentes*. Para esto:

- Se propone analizar el estado del arte en los *Sistemas de Similitud* y *Modelos de Fusión de Datos* para elegir las técnicas más adecuadas con las que se consiga mejores resultados.
- De la revisión bibliográfica, como investigación de la tesis, desarrollar un nuevo *Modelo de Fusión de Datos* que mejore la eficiencia de los diversos *Sistemas de Similitud* que se deseen integrar.

### 1.2.1. Objetivo general

Diseñar y desarrollar un *Modelo de Fusión de Datos* eficiente que pueda integrar diversos resultados de *Sistemas de Similitud entre Documentos*, y que pueda agregar un único valor a las duplas *Documento Sospechoso – Documento Fuente* con mejores similitudes.

### 1.2.2. Objetivos específicos

1. Entender el comportamiento de las técnicas de similitud de documentos, a fin de determinar la probabilidad que un documento tenga grado de copia.
2. Establecer una metodología de estandarización para los resultados de los *Sistemas de Similitud* que entrarán al *Modelo de Fusión de Datos*.
3. Diseñar y desarrollar un *Modelo de Fusión de Datos* que en lo posible incluya valores de calibración parametrizables, para que pueda ser un sistema flexible al cambio.
4. Proponer una metodología de desarrollo para determinar la probabilidad de copia de un documento y mediante juicio experto definir qué documentos presentan plagio, o no.

## 1.3. Resultados Esperados

1. Desarrollar un sistema eficiente de fusión de datos. Que ayude a generalizar los resultados entregados por diversos sistemas de detección de plagio.
2. Comprobar que la metodología de utilizar un modelo de fusión de datos para sistemas de detección de plagio resulta más eficiente en tiempo computacional y en calidad de resultados que los sistemas por separado.
3. Diseñar un esquema de detección de plagio que permita estandarizar el reconocimiento de copia entre documentos escritos.
4. Diseñar un esquema de detección de plagio que sirva como herramienta de apoyo a los análisis de redes sociales (SNA<sup>8</sup>) para reconocer la copia entre un grupo de personas.

---

<sup>8</sup>SNA: Social Network Analysis, por su nombre en inglés

## 1.4. Metodología

Para conseguir los objetivos planteados el proyecto se basará en la definición que describe Armando Asti en [3], donde describe que la metodología adecuada de investigación “*corresponde al estudio analítico de los métodos de investigación y de prueba, incluyendo la descripción de los hechos y su valoración crítica*”. Entonces, organizando el plan, la estructura y la estrategia de investigación se seguirá un lineamiento metodológico dividido en cinco etapas que van desde la problemática inicial hasta la solución y análisis específico:

### 1. El problema del plagio escrito, definiciones y estado del arte

La primera etapa del proyecto de tesis buscará ahondar en definiciones para el problema del plagio escrito, se repasarán algunos conceptos y soluciones algorítmicas que buscan detectar el plagio escrito. Una vez comprendido el comportamiento de las *Técnicas de Similitud*, y sabiendo de antemano que al ser utilizadas por separado no son muy eficientes, se revisará el estado del arte de los *Modelos de Fusión de Datos*.

### 2. Diseño y desarrollo del Modelo de Fusión de Datos

Luego de la revisión bibliográfica, se desarrollará un Modelo matemático para Fusión de Datos. Este modelo deberá integrar los mejores Sistemas para Fusión de Datos que se hayan encontrado en el paso anterior, también tendrá que incluir conceptos que agreguen valor a los resultados analizados. Además, para afinar mejor la ecuación del Modelo de Fusión, se deberán tomar en cuenta valores de calibración del modelo con el fin de mejorar la eficiencia de los diversos Métodos de Similitud integrados.

### 3. Calibración de los parámetros del Modelo

El ajuste de los parámetros del Modelo se realizará por la aplicación de las ecuaciones para *Fusión de Datos* en una base de datos con información supervisada de Documentos Sospechosos y Fuentes. Con el fin de asegurar la mayor efectividad, se propone seguir una estrategia de revisión pseudo-exhaustiva de valores para la calibración de parámetros basado en conceptos del algoritmo Greedy[15].

### 4. Evaluación del modelo de fusión

Esta etapa se dividirá en dos: *Desarrollo de Métodos de Fusión clásicos de la literatura y Comparación entre el Método de Fusión de la tesis y los otros Métodos de Fusión clásicos*. Para la primera parte, se implementarán los Métodos de Fusión de Datos clásicos como los propuestos por Shaw y Fox [63] y Montague [49]. La segunda parte, consiste en comparar los resultados obtenidos con el *Modelo de Fusión de Datos* propuesto y desarrollado en esta tesis versus los otros *Modelos clásicos* con el propósito de verificar la efectividad del nuevo método ante los ya existentes. Los resultados que estos métodos entreguen serán evaluados mediante técnicas de efectividad de predicción en IR (*Precision, Recall y F-measure*)[45, 73].

### 5. Conclusiones

Finalmente, luego de todos los análisis, se darán las observaciones sobre los principales resultados del proyecto. Y se presentarán las conclusiones de este trabajo de acuerdo a los resultados esperados.

## 1.5. Estructura de la tesis

Siguiendo el esquema metodológico definido, el siguiente capítulo de este documento contiene una revisión bibliográfica general de la problemática del plagio y de las técnicas para su detección. Además, se profundiza en el estado del arte de los *Modelos clásicos de fusión de datos*, dando énfasis a las ecuaciones de los algoritmos matemáticos, cuyo entendimiento resulta importante para el posterior desarrollo de la tesis.

En el Capítulo 3, nombrado “Modelo de Fusión de Datos por Score”, se presenta la principal contribución de la tesis, el *Modelo de Fusión de Datos propuesto*. Se describen las novedades y características del modelo propuesto, la lógica de desarrollo, la formulación matemática y su algoritmo de implementación.

El Capítulo “Experimentaciones” (Capítulo 4), presenta la configuración y evaluación del Modelo descrito en el Capítulo 3. Se explica la implementación de los Modelos de Fusión de Datos clásicos abordados en el Capítulo 2 descritos en la literatura [49, 63] con los cuales, apoyado por una base de datos de prueba, se contrastarán los resultados con el modelo propuesto. Además se realizan los cálculos de eficiencia algorítmica para todos los modelos.

El penúltimo capítulo (Capítulo 5) lleva por nombre “Análisis de Resultados” y contiene todos los resultados obtenidos por cada Modelo de Fusión de Datos (clásicos y propuesto) se analizarán los resultados para reconocer al *Modelo* que tuvo el mejor desempeño. En esta misma sección se realiza también un análisis de colusión, para reconocer el grado de copia entre personas de una misma red social.

Finalmente se tiene el Capítulo 6, “Conclusiones”, que después del análisis comparativo del Capítulo 5, presenta las principales conclusiones y contribuciones del proyecto de investigación. Posteriormente se discute el trabajo futuro y las posibles líneas de investigación que esta tesis conlleva.

## Capítulo 2

# Trabajo Previo

En este capítulo, se discute el estado del arte de los tópicos de esta tesis. Primero se presenta el problema del plagio, los distintos estudios que lo abordan y algunos métodos para detección de plagio en documentos escritos existentes en la literatura. Finalmente se abordará el estado del arte en la fusión de datos.

### 2.1. El problema del plagio

El problema del plagio no es un tema aislado, hay muchas implicaciones éticas y penales respecto a él. Sólo un ejemplo del problema se ve en la industria discográfica, donde muchas personas deben pagar millonarias sumas de dinero para resarcir al autor original<sup>1</sup> de la obra. De acuerdo a definición de la Real Academia de la lengua Española (RAE)[27], la palabra *Plagio* proviene del latín *plagiārius* que significa *secuestrar*. Como desvirtuación de ella se desprende la definición actual de plagio que vendría de *secuestrar el trabajo de un tercero para tomarlo como propio*, específicamente la RAE lo define como “*Copiar en lo sustancial obras ajenas, dándolas como propias.*”. Por otro lado, otros autores como Cormeny [16] y Roig [54], citando a la Asociación Americana de Profesores Universitarios<sup>2</sup> coinciden en la definición:

*“Tomar las ideas, métodos o escritos de otra persona sin su consentimiento pero con la intención de engañar”*

Existe evidencia que el problema del plagio se ha incrementado en las últimas décadas por la facilidad de acceso a la información. De acuerdo a Björklud y Wenestam [11]: en Estados Unidos, investigaciones realizadas en instituciones académicas en el año 1941 muestran que el 23 % de estudiantes aceptaron haber

---

<sup>1</sup>A nivel legal se considera *autor intelectual* y esta definición no sólo abarca obras sino también ideas

<sup>2</sup>American Association of University Professor, por su nombre en inglés

copiado de algún modo, mientras que en el año 1952 ese rango se incrementó a 38 %. En el año 1960, desde el trabajo de Donald Thistlethwaite [69] donde cita el estudio de William y Suchman se concluyó que el número de estudiantes que aceptaron haber copiado incrementó a 49 %; cuatro años después, en 1964, Hetherington y Feldman [32] indicaron que ese valor se encontraba en 64 % y en 1980 la investigación de Baird [5] entregó un 76 %. En el año 1995, considerándolo como año de inicio de la masificación del Internet, el estudio llevado por Franklyn-Stokes [30] en Inglaterra entregó que el 60 % de los encuestados admitió haber copiado. De acuerdo a Carroll y Appleton [12] citando el estudio de Walker [75], en 1998, se encontró que la cifra subió al 80 % en sólo tres años desde la explosión de la Internet. En el año 2001, Josephson y Mertz [33]; y McCabe et. al. [47] coincidieron en un valor de 74 %. Para el 2002, Kellogg [34] encontró que alrededor de 90 % de alumnos aceptan haber copiado. Años posteriores, y hasta la fecha, diversos estudios y autores como Coverdale y Henning [17]; Carroll y Appleton [12]; Bordignon et. al. [28]; siguieron concluyendo que el porcentaje de copia no disminuía y que en casos aumentaba. De acuerdo a Comas, Sureda, Ortega y Urbina [57] con el avance tecnológico las malas prácticas de copia han evolucionado.

Una investigación similar llevada a cabo en Chile en el 2008 [14] entregó que el número de personas que copian es superior al 80 %, en el mismo estudio se preguntó si consideraban que al copiar igual se aprende, siendo la respuesta bastante interesante, un 53.6 % de los encuestados afirmó que considera que “*Copiando igual se aprende*”. Los resultados encontrados no deberían sorprender, ya desde 1993 Lipson y McGraven [41] estudiaron que las personas que copian lo hacen porque la copia es percibida como algo muy común, dicho resultado se afirma con el estudio de Davis y Ludvigson [20]. Entonces es un hecho bastante cierto decir que los estudiantes copian, ahora queda la duda “*¿los profesores están capacitados para descubrir la copia?*”, Björklund y Wenestam [11] se encargaron de responder la pregunta, indicaron que sólo un 20 % del cuerpo académico está preparado para detectar la copia. Mencionando el mismo estudio en Chile [14], la percepción de los estudiantes es bastante parecida, un 73 % asume que los profesores si se percatan de la copia. Además, la visión del plagio es que es un delito menor y muchas veces no es sancionado [19], quedando impunes muchos alumnos que lo cometen, por ello no es de sorprender que los alumnos estén dispuestos a “*arriesgarse y copiar*”.

La solución para este problema es utilizar un conjunto de técnicas o métodos para detección de plagio. Estas soluciones permiten descubrir si un documento sospechoso es efectivamente un plagio, es decir fue copiado desde otro documento. El funcionamiento de los sistemas de detección de plagio será presentada más adelante, y ya se dio un breve alcance de su funcionamiento en el Capítulo 1, figura 1.1 (Página 2), pero antes de continuar con la discusión de los métodos de detección de plagio, se entregarán algunas definiciones previas.

## 2.2. Tipos de plagio

Como ya se ha discutido, y de acuerdo al estudio de Fly [29], el plagio se ha ido incrementando con el avance de las tecnologías que dan acceso a la información, como por ejemplo la Internet, radio y televisión. Si el objetivo es detectar el plagio e identificar quienes lo cometen, es esencial que se conozca sus tipos y variaciones las cuales se pueden distinguir:



- Por la intención:
  - **Copia con intención deliberada.** Cuando el objetivo del trabajo es hacerlo copiando.
  - **Copia con intención directa.** Cuando el trabajo presenta algunas partes, párrafos o frases, copiadas textualmente y no citadas.
  - **Copia sin intención.** Cuando el autor no tiene el conocimiento de estar incurriendo en el delito de plagio.
- Por la estructura:
  - **Copia textual.** Cuando el documento sospechoso, o un fragmento de este, es idéntico al documento original.
  - **Copia con variaciones estructurales.** Cuando el documento sospechoso presenta cambios textuales respecto al original.
  - **Copia con variaciones gramaticales.** Cuando el documento sospechoso fue traducido porque el original es de idioma distinto.
- Por el origen de la fuente:
  - **Copia externa.** Cuando el documento sospechoso fue copiado desde una fuente externa.
  - **Copia interna.** Cuando el documento sospechoso fue copiado desde sí mismo.

A continuación se realizará una breve descripción de cada tipo.

### 2.2.1. Copia por intención

En este tipo se asume que la copia existe y lo que se analiza es la decisión del autor del documento copiado: Si tuvo la intención directa de copiar o si fue una copia no intencional.

- **Copia con intención deliberada**

La copia deliberada concentra en gran medida la definición del plagio, como lo indica un artículo de “prevención del plagio”<sup>3</sup> publicado en 2005 por la Universidad de Northwestern [68] “*La copia deliberada es el utilizar el trabajo de otros y convertirlo como el trabajo propio*”. Esta copia puede ser de diversas fuentes: Ensayos, resúmenes, artículos enciclopédicos, libros, revistas, papers, entre otros. No se es necesario profundizar en definiciones, este tipo de plagio es totalmente perceptible y es una copia directa se le analice por donde se le analice, Mario Nuñez en el 2006 [52] presentó una serie de pautas para reconocer la copia con intención deliberada. Es copia intencional cuando:

- Se compra, se roba o se toma prestado un trabajo redactado por otra persona para hacerlo pasar como propio.

---

<sup>3</sup>Avoiding plagiarism, por su nombre original en inglés

- Se le paga a otra persona para que escriba el trabajo que se hará pasar como propio.
- Se copian a drede las ideas de otros, sin darle crédito, para hacerlas pasar como propias.

Este tipo de plagio no tiene puntos, ni defensa, a su favor. El delito del plagio se comete intencionalmente y todo lo evidencia, además el autor de la copia no desarrolla ningún tipo de habilidad.

#### ■ Copia con intención directa

Este tipo de copia también es intencional pero, a diferencia de la *copia con intención deliberada*, para ser descubierta se requiere de un esfuerzo mayor de análisis del documento sospechoso. Es decir, el autor de la copia tiene la intención de copiar pero para evitar ser descubierto se encarga de trabajar más el documento de copia, debido a esta consideración es que la detección de copia se vuelve difusa porque no siempre se puede afirmar que el documento sospechoso sea copiado. George MacDonald en su publicación de [44] indica que para este tipo de copia el parafraseo es lo más usual, el cual consiste en cambiar algunas palabras, o frases, por otras pero manteniendo la misma idea del documento original. Esta práctica suele ser tan aplicada que MacDonald cita una pagina web de la Universidad de Indiana en Estados Unidos<sup>4</sup> donde explican el modo en *cómo no cometer plagio*, y muestran ejemplo de parafraseos considerados *aceptables* y *no aceptables*. El argumento para considerar de *no aceptable* a un parafraseo es que “*Sólo pocas palabras o frases hayan sido cambiadas*”. Esto marca una tendencia de pensamiento interesante:

- Cuando se realiza un *parafraseo superficial*, el documento se considera como una copia porque el autor sólo se limitó a copiar la misma idea y hacer muy pocos cambios.
- Cuando se realiza un *parafraseo profundo*, el documento se considera nuevo porque el documento ya no es el mismo, además el autor del parafraseo está logrando capacidad de síntesis y comprensión del tema.

Sea cual sea la mentalidad, en el parafraseo se sigue cometiendo delito de plagio si no se citan las respectivas fuentes. Si bien la capacidad para resumir lo que algún otro autor dijo en sus propias palabras es buena, es mucho mejor la capacidad de citar a otro autor, analizar la idea que dicho autor quiso entregar y luego expresar el concepto dentro del nuevo contexto del documento. En este nivel es donde se demuestra que se ha desarrollado la capacidad de pensar por si mismo y que se es dueño de un documento original.

#### ■ Copia sin intención

Debido a la discusión que *resulta absurdo esperar que los autores hagan referencia de cada cita que agregan a sus documentos* es que puede surgir este tipo de copia. La *copia sin intención* ocurre cuando, por diversos factores, el autor no coloca referencias a un cita o investigación realizada por algún tercero dando a entender que lo descrito es algo propio. Nuñez en [52] indica que la copia sin intención ocurre por tres factores:

- Falta de conocimiento.

Este caso es cuando el autor no sabe citar, o no cita correctamente sus fuentes y, por lo tanto, no tiene consciencia de estar cometiendo delito de plagio. Este caso, suele ocurrir en autores novatos

<sup>4</sup>Hoy en día el mencionado vínculo web ya no existe (<http://www.indiana.edu/~wts/plagiarism.html>), el estudio de MacDonald fue realizado en el año 2003.

o poco experimentados en la redacción porque ellos aún no han adquirido todo el conocimiento para hacer un documento formal. Este caso, en el ámbito académico, considera que los educadores están fallando en inculcar los conocimientos básicos sobre redacción de documentos a sus estudiantes.

- Problema del parafraseo.

Haciendo referencia a la copia con intención directa, donde los educadores ven como una buena práctica el parafraseo profundo sin referenciar el origen del documento, muchos autores cometen el delito del plagio porque siguen las pautas aprendidas. El documento seguirá siendo una copia, así todas las palabras hayan sido cambiadas, o parafraseadas, siempre que se siga con la misma idea original en la estructura original y no se referencie el origen de la fuente.

- Influencia de ideas.

La copia, en este caso, se da porque el autor del documento se encuentra influenciado por las ideas de algún tercero, de modo que su documento final contiene dichas ideas, pero el autor no da los respectivos créditos a la persona que tuvo la idea originalmente porque cree que aquella línea de pensamiento es propia y no posee influencias. Este caso se suele dar en documentos donde las ideologías o los conceptos subjetivos son muy importantes, por ejemplo en las áreas académicas sociales (psicología, la política, sociología, entre otros).

La copia sin intención se puede considerar como el tipo de copia menos nocivo porque, si bien el autor del documento no tiene consciencia que está incurriendo en el delito de plagio, si está dedicándose en redactar un documento original y de ese modo demostrando que tiene desarrollada la capacidad de pensar y plasmar sus propias ideas o estructura, con la carencia que no ha sabido referenciar adecuadamente su trabajo.

### 2.2.2. Copia por estructura

El tipo de copia por estructura, como su nombre lo indica, describe la disposición y orden de las partes del documento original dentro del documento copiado. Eissen y Stein en [26] hacen mención de lo que denominan la taxonomía del plagio (Figura 2.1), donde clasifican a este tipo de copia de acuerdo a las características de los documentos copiados.

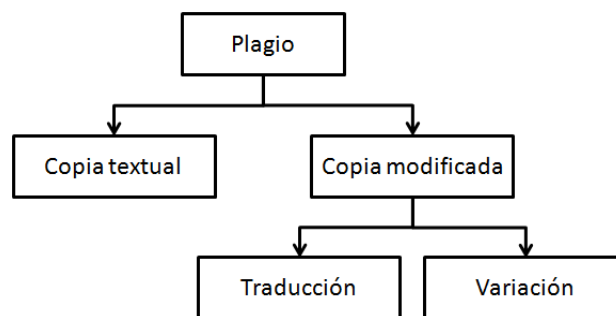


Figura 2.1: Taxonomía del plagio [26]

En la figura 2.1, se puede observar que la copia por estructura tiene subdivisiones que son:

■ **Copia textual**

Este caso de copia se da cuando el texto original es copiado íntegramente sin cambios de ningún tipo. La estructura del documento copia contiene el mismo texto, estilo de escritura y distribución de párrafos idéntico al documento original. Si el documento copia fue alterado en el formato de letra (tamaño y estilo de letra) también se considera copia textual.

■ **Copia con variaciones estructurales**

Este tipo de copia también suele ser denominado *copia* semántica. La copia con variaciones estructurales ocurre cuando el texto copia posee párrafos, frases u oraciones del texto original que han sido modificados con el objetivo de *enmascarar* la copia textual, esta definición es equivalente a la definición del *parafraseo* entregado en el punto 2.2.1 de la página 12.

De acuerdo al estudio realizado por Comas et. al. [57], con el pasar de los años académicos los estudiantes que tienen el hábito de copiar han ido evolucionando sus técnicas de plagio, realizando cada vez mayores y más complejas variaciones estructurales como:

- Reemplazo por sinónimos, palabras rebuscadas<sup>5</sup> por otras más usuales, o viceversa.
- Contenido cambiado agregando las denominadas *palabras de relleno*.
- Cambio de títulos o titulares, pero manteniendo la misma idea.
- Eliminando secciones, o recortando párrafos muy grandes, pero sin cambiar fuertemente la idea o el contenido.
- Re-ordenando secciones o párrafos dentro del documento.

■ **Copia con variaciones gramaticales**

La copia con variaciones gramaticales sucede cuando el documento copia se realiza desde un documento original que está escrito en otro idioma y este es traducido. Como toda traducción textual no es posible, el texto copia debe ser modificado para que el escrito tenga coherencia en el nuevo idioma. De acuerdo a la revista electrónica Información Electrónica<sup>6</sup> [59] este tipo de copia ha logrado presencia entre un círculo de investigadores con poca ética que utilizan investigaciones de terceros que publican en idiomas poco conocidos internacionalmente (urdú, mongol, birmano, nepalés, tibetano, etc.) los traducen al idioma inglés y luego los presentan como propios en prestigiosas revistas.

### 2.2.3. Copia por el origen de la fuente

El tipo de copia por el origen de la fuente entrega información desde dónde se extrajo la información para el documento sospechoso. Si bien parece bastante natural que la copia se haga desde alguna fuente externa, existen casos donde la copia se realiza desde la misma fuente. Por ello se definen los sub-tipos de Copias por el origen de la fuente:

<sup>5</sup>Se denominan así a las palabras que no son del dialecto común del entorno del autor o de la época.

<sup>6</sup>Revista distribuida electrónicamente

■ **Copia externa**

Copia de una fuente externa (libro, artículo, página web, revista, etc.) o también denominado *plagio externo*<sup>7</sup>. Este es el tipo de copia más común, cuando un autor toma como propias las ideas de terceros. No se hace necesario ahondar en más detalles de este tipo de copia porque en las secciones anteriores ya se hizo una amplia revisión del tema en cuestión.

■ **Copia interna**

Copia desde el mismo documento, también se le denomina *plagio intrínseco*<sup>8</sup>, usualmente se da cuando hay más de un autor en el documento [6, 60]. Por ejemplo, en un libro de cierta área temática que contiene trabajos de diversos autores, uno de los autores escribe acerca de un tema que él menciona como propio, más adelante en el mismo documento un segundo autor con su propio estilo y con sus propias palabras escribe del mismo tema también argumentando como suyo. Si bien este tipo de copia no siempre es directa, porque se puede relacionar con la *influencia de ideas* en la sección 2.2.1 de la página 12, igual está considerado como una clase de copia.

**2.2.4. Colusión**

La colusión no es exactamente otro tipo de plagio, por ello no puede ser clasificada directamente dentro del grupo de *tipos de copia* sino que se refiere a un hecho a nivel social donde un grupo de personas realizan un convenio hecho de forma fraudulenta y secreta, con el objeto de cometer plagio entre ellos. De acuerdo a la definición de Jenny Moon [50] la colusión significa tomar el trabajo de otra persona y hacerlo pasar como propio, con la diferencia que la persona de quien se está copiando está consciente del plagio. En otras palabras, es el acuerdo común entre el autor original y el autor que copia para que se realice el plagio de documentos.

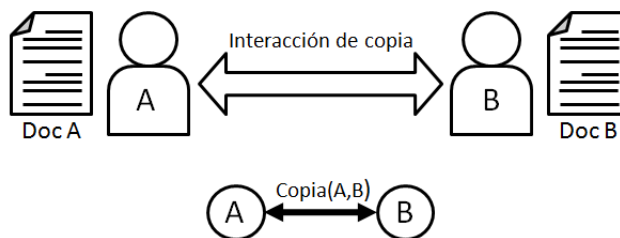


Figura 2.2: Red Social de Copia - 2 individuos

Este tipo de casos se suele dar en el campo académico, donde un grupo de *amigos o conocidos* toma el acuerdo común de “ayudarse” con la entrega de trabajos escritos. Esta interacción entre individuos crea las denominadas redes sociales de copia, la cual ha significado mérito de diversos estudios [2, 22, 71]. En la Figura 2.2 se muestra el acuerdo entre dos individuos para interactuar entre sí y copiar sus documentos, esta interacción puede ser modelada matemáticamente mediante teoría de grafos  $G = (V, E)$ , como un grafo dirigido donde los documentos de los individuos son los nodos ( $V$ ) y el grado de copia entre sus

<sup>7</sup>Extra-corporal plagiarism, en inglés

<sup>8</sup>Intra-corporal plagiarism, en inglés

documentos son los pesos  $C(A, B)$  de los arcos ( $E$ ). La interacción entre grupos sociales de copia puede ser mayor, como se muestra en el ejemplo de la Figura 2.3.

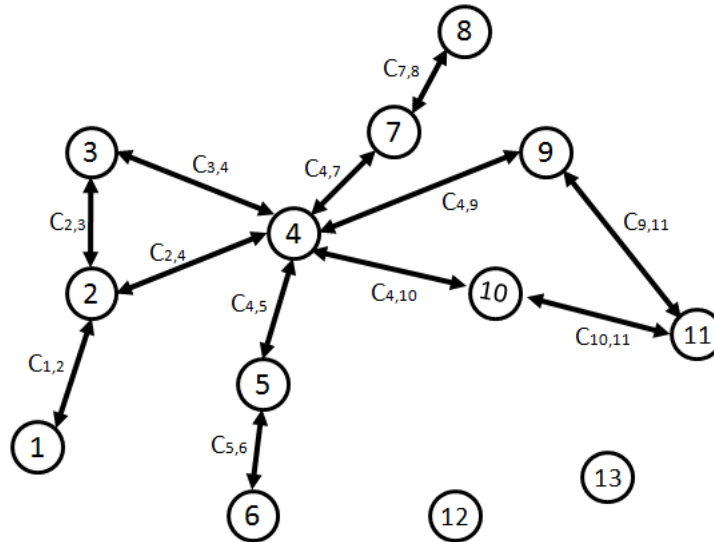


Figura 2.3: Red Social de Copia - 13 individuos

Donde:

$$c_{i,j} = c_{j,i} \quad ; \quad i \neq j, \forall i \in V, \forall j \in V$$

En este caso, se nota una diversidad de colusiones donde el individuo se relacionó con:

- 1 persona: Nodos 1, 6, 8
- 2 personas: Nodos 3, 5, 7, 9, 10, 11
- 3 o más personas: Nodos 2, 4,

Además se distinguen los nodos 12 y 13 que no tuvieron interacción con ningún otro individuo, por lo tanto sus documentos no fueron copiados de la red  $G = (V, E)$ ; incluso se podría llegar a afirmar que son trabajos originales.

### 2.3. Métodos para detección de plagio

De acuerdo a la investigación de Brian Martin [46], la mayoría de documentos copiados resulta ser de intención deliberada, tipo copia textual y externa. En otras palabras es una copia simple y directa por lo tanto, si la fuente es única y conocida, su detección no resulta difícil. Pero en los casos que la copia

procede de diversas fuentes, posee cambios de estructura o ambos; la detección de copia se hace compleja y ocasionalmente imposible de detectar. Entonces se debe recurrir a técnicas para detectar la copia entre documentos.

### 2.3.1. Métodos manuales de detección de plagio

Los educadores, que a su vez son evaluadores de documentos poseen diversos métodos para reconocer si un documento sospechoso es copia de algún otro. Peggy Bates y Margaret Fain de la Universidad Coastal de Carolina en su publicación “*Detectando copia*”<sup>9</sup>[7] hacen una recopilación de diversos métodos manuales para reconocer copia entre documentos:

- Estilo de ciertos párrafos diferentes al resto del documento: Escritura, lenguaje, vocabulario, tono, gramática, etc.
- Formas de plasmar las ideas con palabras que no pertenecen al dialecto nativo del autor. Por ejemplo: Uso de vocablos o jergas del idioma argentino cuando el autor es mexicano.
- Secciones del documento que no poseen relación con el resto del escrito. En otras palabras, el *documento original* no tiene relación con el tema del *documento copia* por ello existen *modificaciones* realizadas por el autor modificar con el objetivo de *fusionar* los temas.
- Numeración de páginas no consistente, esto indica que el documento copia fue tomado de un documento original que posee un mayor número de páginas.
- El uso de pronombres que hacen referencia al autor del documento no corresponden al género del autor.
- Existe texto extraño en el encabezado y/o pie de página del documento copia, como por ejemplo numeraciones que no corresponden, direcciones web y nombres de editoriales. Esto indica que el documento tiene otro origen (web, fotocopia, etc.)
- Documento copia con marcas de agua que indican la procedencia del escrito desde alguna fuente externa.
- Referencia a gráficos, tablas, esquemas o material adicional que no se encuentra en el documento.
- Referencia a personas que participaron directamente en la elaboración del documento que el autor no conoce.
- El documento contiene frases referenciadas que no están incluidas en la fuente citada. Es decir, un documento con una ecuación matemática referenciada apunta a un libro de psicología.
- Los sitios web referenciados como fuentes del documento, están inactivos desde antes que el autor inicie la redacción.

---

<sup>9</sup>Detecting Plagiarization, en su nombre original

- Todas las referencias son de materiales de alta antigüedad. Un valor práctico es encontrar que sean de 5 o más años de antigüedad para reconocer que hay algo *extraño* en el documento.
- El documento posee frases que indican temporalidad de eventos como actuales, cuando en el momento de la redacción ya eran hechos pasados.
- En una confrontación con el autor: Cuando se le pide que pruebe sus fuentes, este no lo puede hacer.
- En una confrontación con el autor: El autor no puede resumir los puntos principales de su documento, ni responder a preguntas específicas del mismo
- En una confrontación con el autor: Si se extrae una frase o párrafo del documento, se le eliminan algunas palabras y luego se le presenta al autor para que las complete, este no es capaz de encontrar la palabra exacta, ni un sinónimo razonable

### 2.3.2. Métodos numéricos para detección de plagio

El objetivo de los métodos numéricos para detección de plagio, es automatizar la lógica humana para reconocimiento de plagio entre documentos. Sin hacer una descripción exhaustiva se presentan algunos métodos representativos para detección de plagio.

- **Coincidencia de cadenas de caracteres.**

Este método, en la definición de Dan Gusfield [31], busca identificar el número máximo de coincidencias entre un par de cadenas de caracteres. Usualmente las cadenas de caracteres se representan como *árboles de sufijos*<sup>10</sup> y se emplean conceptos de teoría de grafos para detectar la cadena de caracteres con mayor similitud entre ellas.

- **Similitud de tópicos**

La lógica de similitud de tópicos consiste en analizar el documento por párrafos y, mediante una lógica de pesos de palabras, extraer los tópicos más representativos de cada párrafo. Posteriormente, mediante comparación de tópicos entre documentos, se calcula la similitud de documentos y así se reconoce el plagio. Este método ha sido abordado en los trabajos de Si, Leong y Lau [65]; Bernstein y Zobel [10], donde hacen la comparación de tópicos entre documentos, si la similitud excede cierto *threshold* los documentos fuente y sospechoso son sub-divididos en partes más pequeñas que un párrafo (usualmente oraciones) y luego son comparados nuevamente. Este proceso puede ser recursivo y llegar hasta la unidad mínima de copia del documento<sup>11</sup>

- **Comparación de n-gramas**

La idea de la comparación de N-gramas es similar a la *Coincidencia de cadenas de caracteres* en el sentido que buscan similitud entre un grupo de caracteres. En este caso en particular, el método

<sup>10</sup>Suffix trees, por su nombre en inglés

<sup>11</sup>Unidad mínima de copia (UMC): Es el mínimo número de palabras que pueden ser agrupados para que al ser analizados contra otros UMC por un método de similitud de copia se pueda reconocer adecuadamente que tan parecidos son entre ellos



agrupa un determinado número de n-palabras (n-gramas) de un documento y las compara con otros n-gramas de otro documento; si encuentra similitud entre ellos les agrega un peso relativo a la copia entre documentos. Trabajos como el de Oberreuter [53] en el 2010, utilizan esta lógica de n-gramas adicionando otros factores como el denominado uso de *ventanas*: la comparación por n-gramas por documento se realiza en un conjunto de muchos gramas, denominados *ventanas*, siendo la similitud entre n-gramas un factor que depende de la ventana. Es decir, mientras más conjuntos de gramas similares se encuentren en una ventana más peso de similitud tendrá el análisis entre documentos.

### 2.3.3. Software para detección de plagio

En el ámbito comercial, existen compañías que ofrecen sistemas para detección de copia automática utilizando diversos métodos [70]. A continuación se detallan algunos softwares para detección de copia:

#### 1. DOC Cop plagiarism detection (<http://www.doccop.com/><sup>12</sup>)

Este es un detector de plagio de origen australiano, creado por Mark McCrohon, gratuito, trabaja con el idioma inglés y está conformado por una página web donde se realiza el análisis de similitud entre documentos vía on-line. Para utilizarlo es necesario contar con una ID de usuario, la cual se obtiene mediante una suscripción en la página web de DOC Cop.

El software admite dos tipos de análisis:

- **File check.** Comparación entre documentos del computador del usuario. Las características de los documentos a analizar es que sean archivos con extensión doc, docx o pdf. El número máximo de documentos a compararse son 8 y el número de palabras de los documentos debe encontrarse entre 20 y 100,000.
- **Web check.** Comparación entre un texto ingresado y la web. El máximo número de caracteres ingresados en la caja de texto es de 888.

Cualquiera sea el análisis utilizado, DOC Cop entrega un reporte vía correo electrónico en formato html, doc o docx según lo elija el usuario. El tiempo de procesamiento, de acuerdo a la página web oficial, es de pocos minutos a un máximo de una hora. Otra característica importante del sistema es que, de acuerdo a sus políticas y términos de uso, DOC Cop no se queda con una copia en sus servidores del documento analizado; de modo que no pueden haber problemas futuros por la privacidad del documento o actos ilegales respecto a la información contenida en los documentos.

#### 2. Docoloc (<http://www.docoloc.de/><sup>13</sup>)

Docoloc es un sistema de origen Alemán, creado por *Institut für angewandte Lerntechnologie (IFALT)*, no es gratuito pero posee una versión demo de características limitadas, trabaja para el idioma inglés y alemán; y posee una interfaz para análisis on-line de los documentos. Para utilizarlo es necesario contar con una ID de usuario que se obtiene en la página web oficial del producto, y además es requerida una

<sup>12</sup>Último acceso: 14-Sep-2010

<sup>13</sup>Último acceso: 14-Sep-2010

licencia de uso para desbloquear las opciones de análisis del sistema. De acuerdo a la información oficial de la página web, el sistema admite archivos con extensiones PDF, DOC, RTF, HTML, PPT, XLS y texto plano (txt). Docoloc realiza un análisis entre el documento subido por el usuario y la web. Utiliza la misma lógica que un motor de búsqueda web, recorre e indexa la web mediante un crawler almacenando la información de las páginas web visitadas en un servidor. Posteriormente extrae esta información para compararla con el archivo subido por el usuario. El tiempo de procesamiento es relativo al tamaño del documento a analizar. El reporte final puede ser visualizado de 3 modos distintos: Un envío al **Correo electrónico.**, en la página de acceso personal o ambas opciones anteriores. Los documentos ingresados se quedan en el servidor de Docoloc para fines de respaldo de la información, de acuerdo a sus términos de uso.

3. **Ephorus** (<http://www.ephorus.nl/><sup>14</sup>)

Sistema de origen Holandés, cuyo nombre proviene del griego que significa *inspector de enseñanza*, fue creado por el grupo Ephorus B.V., es un detector de plagio de pago pero permite el uso de una demo con características limitadas; el sistema es independiente del lenguaje siempre que se escriba en texto con letras alfabéticas; es una interfaz simple y on-line. Se requiere de un ID y contraseña para utilizarlo. De acuerdo a la información de su página web, el reporte de información es simple. Posee la función de comparar un documento sospechoso contra un grupo de documentos fuentes, creación de una base de datos particular que puede ser utilizada para encontrar plagio en futuros procesos. Además, guarda automáticamente los documentos de análisis en su base de datos de acuerdo a sus términos de uso.

4. **Essay Verification Engine (EVE2)** (<http://www.canexus.com/><sup>15</sup>)

EVE2 fue creado por la compañía Canexus; a diferencia de los otros sistemas este no es un detector de plagio on-line sino que consta de un software ejecutable que se instala en la PC del usuario; este software es de pago y funciona con licencia, sin embargo permite un uso libre hasta por 15 días desde la instalación; trabaja adecuadamente para el idioma inglés logrando una efectividad alrededor del 90 % en encontrar el material copiado de acuerdo a la página oficial. El software acepta documentos en texto plano y archivos de word (extensión doc). El sistema hace uso de la Internet para revisar si el documento sospechoso ha sido copiado de la Web. No hay información si EVE 2 almacena el documento en una base de datos propia sin consentimiento del usuario.

5. **Plagiarism Checker.com** (<http://www.plagiarismchecker.com/><sup>16</sup>)

Este es un sistema gratuito, creado por Darren Hom, un profesor de California, Estados Unidos; Cuya motivación era disuadir a sus alumnos de cometer prácticas de plagio. Respecto al funcionamiento del sistema, Plagiarism Checker.com es un detector de plagio básico que usa los motores de búsqueda Google y Yahoo para identificar el origen de una frase que el usuario ingresó manualmente desde algún trabajo. Los resultados obtenidos de las búsquedas indicarán que tanta copia textual hay en un documento, o no. Cabe resaltar que los resultados se muestran en las páginas oficiales de los buscadores y las opciones por defecto de cada motor de búsqueda aplican directamente. Por ejemplo, colocar entre comillas una frase para buscar una coincidencia exacta en Google. Al ser un sistema básico, solo se

---

<sup>14</sup>Último acceso: 14-Sep-2010

<sup>15</sup>Último acceso: 14-Sep-2010

<sup>16</sup>Último acceso: 14-Sep-2010

encarga de reconocer copias textuales y como utiliza la tecnología de Google y Yahoo, está delimitado a sus idiomas y la frase ingresada manualmente no debe exceder a las admitidas por los buscadores.

6. **Catchitfirst.com** (<http://www.catchitfirst.com/><sup>17</sup>)

Catchitfirst es un detector de plagio on-line, no gratuito, creado en Canadá por Vancouver Software Labs Inc. (VSL). Para ingresar al sistema es necesario contar con un ID y contraseña, que se obtiene previo registro en la página web y pago de la membresía. La comparación de los documentos sospechosos se realiza contra toda la web. Todos los documentos analizados se quedan almacenados en la base de datos de *Catchitfirst* y los reportes se muestran vía on-line. El *Catchitfirst* fue desarrollado a partir de un proyecto previo para detección de plagio de VSL llamado *Scriptum* (<http://www.scriptum.ca/>), la página aún contiene información del funcionamiento hasta un video instructivo pero el sistema está deshabilitado y de acuerdo a la información de sitio, la técnica de detección de plagio fue mejorada e incluida en el *Catchitfirst*.

7. **TurnItIn** (<http://turnitin.com/><sup>18</sup>)

El sistema TurnItIn es un sistema no gratuito para detección de plagio, de los más usados por sus funcionalidades que además de detectar el plagio permite a los estudiantes revisar de manera on-line sus trabajos con el profesor de modo que permite una retroalimentación y aprendizaje. Fue desarrollado por la compañía *iParadigms, LLC* en California, Estados Unidos. TurnItIn posee un modo de pago distinto a los anteriores sistemas, porque como mínimo uno debe comprar la licencia para todo un departamento, sección o escuela, y no para un usuario particular. El funcionamiento de este sistema permite buscar vía on-line el documento sospechoso o compararlo contra una propia base de datos. Los reportes se entregan vía on-line, mediante información visual respecto al porcentaje de copia que presenta una sección del documento sospechoso versus alguna fuente. Además, TurnItIn se queda con una copia en su base de datos del documento analizado.

8. **Urkund** (<http://www.orkund.com/><sup>19</sup>)

El detector de plagio Urkund fue desarrollado por una compañía Suiza llamada PrioInfo AB en el año 2000, por la iniciativa de un grupo de profesores. Funciona como un sistema on-line, con la diferencia que el envío de trabajos no se hace mediante el ingreso a una cuenta de alumno (para subir el trabajo) o de profesor (para analizar los trabajos) sino que Urkund provee de una dirección de correo electrónico a la cual los alumnos deben enviar sus trabajos; y luego del proceso Urkund envía un e-mail al profesor con un vínculo en donde puede encontrar todos los trabajos enviados por los alumnos y el reporte de plagio vía una interfaz web. El sistema es de pago y el análisis sólo lo realiza para el idioma inglés y está más orientado a papers científicos y trabajos académicos de estudios superiores. En la página web oficial existen un vínculo para acceder a una demo del sistema.

Como detalle adicional, la compañía PrioInfo ha establecido contactos con diversas entidades internacionales para que distribuyan oficialmente su producto, por ello a Urkund se le puede encontrar como producto de las compañías *ebrary, Knovel, LexisNexis, Thomson-Gale* y *OvidPrioInfo*.

---

<sup>17</sup>Último acceso: 14-Sep-2010

<sup>18</sup>Último acceso: 14-Sep-2010

<sup>19</sup>Último acceso: 14-Sep-2010

9. **WCOPYFIND** (<http://plagiarism.phys.virginia.edu/Wsoftware.html><sup>20</sup>)

El WCopyfind es un programa gratuito para detección de plagio creado por el profesor Louis Bloomfield de la Universidad de Virginia en Estados Unidos. Este sistema consta de un software ejecutable que se puede descargar desde la página web oficial y que se instala en el computador del usuario. Su funcionamiento se limita al análisis local de documentos, es decir una búsqueda en el computador donde se encuentra instalado el programa, y no hace una búsqueda de documentos en la Internet.

El software permite encontrar documentos que tienen mucha similitud entre el texto contenido. Reconoce copia textual o con muy pequeños cambios. Wcopyfind analiza archivos de texto plano (txt) y documentos de word (doc).

10. **GLATT PLAGIARISM SCREENING PROGRAM** (<http://www.plagiarism.com/><sup>21</sup>)

Glatt plagiarism no es exactamente un sistema de detección automático como todos los anteriores. A diferencia del resto, no utiliza algún método de similitud y detección de plagio que le permite auto-identificar el grado de copia entre documentos; sino que consta de una serie de métodos y recomendaciones para que los profesores puedan identificar el plagio, además puedan aplicar técnicas de escritura con el fin de evitar el plagio. Una de las técnicas para evitar el plagio que utiliza Glatt, es que toma el texto plano del trabajo digitalizado del alumno y aleatoriamente coloca cuadros en blanco sobre algunas palabras. Entonces se confronta con el autor del documento y se le solicita que complete las palabras faltantes, se supone que si el autor conoce su documento podrá colocar la misma palabra o sinónimos cercanos en dicho espacio. Esto es similar a una técnica de detección de plagio manual que se explicó en la sección 2.3.1 página 16.

11. **DOCODE: DOCUMENT COPY DETECTOR** (<http://www.docode.cl/><sup>22</sup>)

El sistema DOCODE se divide en dos tipos: la versión lite que es de acceso libre y la versión profesional que es de pago. La versión lite consiste en un buscador de similitud de copia y entrega, por orden de relevancia, todos los vínculos web que tienen cercanía con el texto ingresado. Su funcionamiento es bastante similar al de cualquier buscador, con la diferencia que no tiene un límite máximo de palabras en el cuadro de búsqueda y no se limita a un único buscador. La versión profesional permite la búsqueda de un documento sospechoso de un modo más avanzado: permite cargar todo el archivo, auto-reconoce la extensión, parsea el texto y luego utiliza el mismo buscador de la versión lite para encontrar en la web todos los posibles documentos fuentes. Posteriormente, mediante diversos métodos de detección de plagio, reconoce el nivel de copia entre la dupla Documento Sospechoso - Documento Fuente; finalmente el proceso entrega un puntaje que indica el porcentaje de similitud de la dupla, dejando a criterio del usuario definir si se trata de un plagio o no.

---

<sup>20</sup>Último acceso: 14-Sep-2010

<sup>21</sup>Último acceso: 14-Sep-2010

<sup>22</sup>Último acceso: 21-Ene-2011

## 2.4. Métodos de fusión de datos

Como se detalló en los objetivos (Sección 1.2), para este proyecto de tesis se desarrollará un Modelo de Fusión de Datos, los datos que fusionará el sistema no son otros que los puntajes del nivel de copia que se obtienen de sistemas individuales para detección de plagio.

Los Sistemas de Detección de Plagio indican la similitud existente entre un documento A y un documento B (ver figura 1.1 en la página 2). En la realidad, se cuenta con un documento sospechoso que debe ser evaluado contra un conjunto de  $n$  documentos fuentes (Ver figura 2.4), como el sistema de detección de plagio entrega un valor por cada comparación “sospechoso–fuente”, entonces se tendrán muchos resultados que ordenadamente se pueden visualizar como un vector de longitud  $n$ .

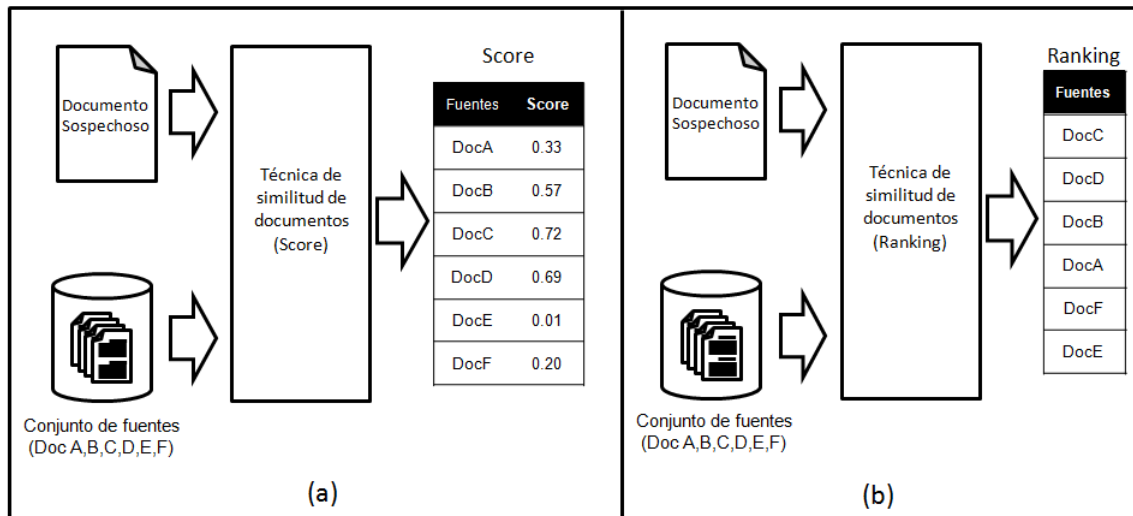


Figura 2.4: Sistema de similitud de documentos (doc. sospechoso vs. docs. fuentes)

En el **caso (a)** el elemento  $k$ -ésimo del vector será el puntaje (score) resultante de evaluar el documento sospechoso contra el  $k$ -ésimo documento fuente. (Donde  $k = 1, 2, 3, \dots, n$ ), en este caso los resultados se encuentran en el rango entre 0 y 1 ( $score \in [0, 1]$ ), sin embargo el rango puede ser cualquier valor real ( $score \in [a, b]; a, b \in \mathbb{R}$ ). En el **caso (b)** los elementos del vector indicarán la similitud con el documento sospechoso; en ese contexto el primer elemento representará al documento fuente con más similitud con el documento sospechoso, en el caso de la figura 2.4 el *documento C* será el más similar al *documento fuente*, y le seguirá el *documento D*, *documento B*, etc.

Si bien los métodos de similitud para los casos (a) y (b) aunque parecen excluyentes uno del otro (Ver figura 2.4), en realidad son muy cercanos, como se observa en la figura 2.5, el *método de detección de similitud con ranking* se origina como consecuencia del *método de detección de similitud con scores* a partir de un ordenamiento ascendente de los resultados entregados por el segundo método; usualmente este ordenamiento ascendente se realiza internamente por el *método de ranqueo* y en consecuencia no se pudo conocer el score inicial asignado a cada documento fuente.

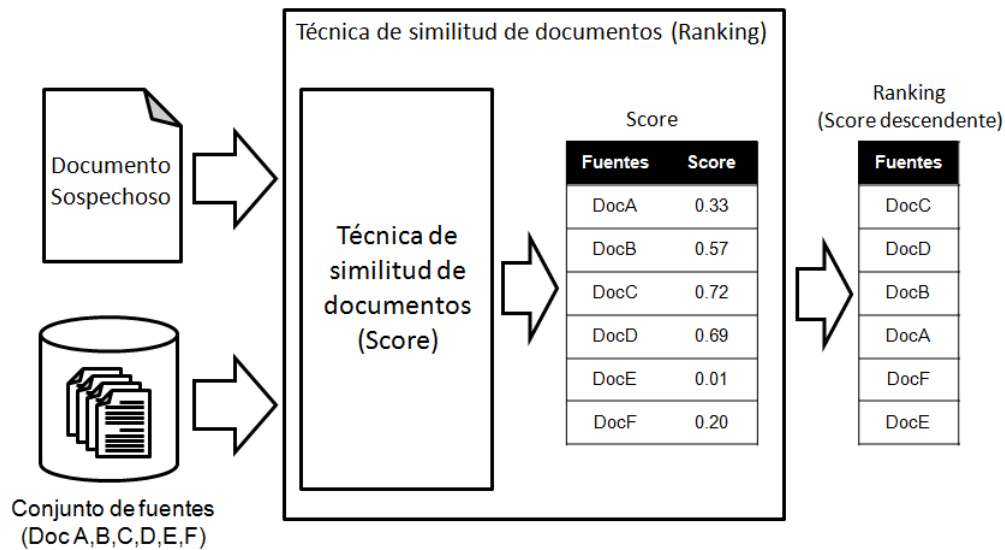


Figura 2.5: Relación entre los métodos de similitud de *Score* y de *Ranking*

Un detalle a considerar para los métodos de fusión de datos, es la elección de los métodos de similitud de documentos. Todos los sistemas de similitud deben contener el mismo tipo de salida, es decir sólo se podrán fusionar entre sí métodos del tipo *Score* o sólo métodos del tipo *ranking*, pero no una mezcla de ambos. En la figura 2.6, se muestra una descripción genérica de un método de fusión de datos para métodos de similitud por score (aunque todo el proceso aplica del mismo modo para ranking). Se cuenta con un documento sospechoso y  $n$  documentos fuentes en una base de datos, además se tienen  $m$  métodos de similitud de documentos ( $m$  no necesariamente es igual a  $n$ ). De acuerdo a lo explicado en la figura 2.4 cada  $k$ -ésimo método de similitud de documentos (donde  $k = 1, 2, 3, \dots, m$ ) entregará resultados del documento sospechoso versus cada uno de los  $n$  documentos fuentes en un vector de  $n$  elementos. En este punto recién inicia el método de fusión de datos, los  $m$  vectores con  $n$  resultados que entregan los métodos de similitud ingresan al sistema de fusión de datos el cual se encarga de integrar y resumir los  $m$  vectores en un único vector respuesta que contiene  $n$  resultados, uno por cada documento fuente existente en la base de datos. Finalmente, este vector indicará el score o ranking que resume los resultados de todos los  $m$  métodos de detección de similitud de documentos.

Para fundamentar el uso de un método de fusión de datos y como ya se ha discutido en las subsecciones anteriores, los métodos de detección de similitud pueden ser diversos: Método manuales, numéricos y sistemas que integran los dos anteriores. Si bien ayudan a reconocer la existencia de copia entre documentos, en la práctica ningún método de detección es perfecto en el reconocimiento de plagio. El error en el reconocimiento sucede por el diseño y construcción de cada método de detección de plagio, pues se les elabora y contrasta para un determinado grupo de documentos donde funcionan adecuadamente otorgando una gran certeza de detección, pero cuando se les evalúa con un universo mayor de documentos (documentos que contienen características nuevas en estructura, formato, estilo, etc.) muchos métodos empiezan a fallar y su certeza disminuye. Por ello, se propone realizar un sistema que integre diversos métodos de detección de plagio, de modo que exista complementariedad entre ellos.

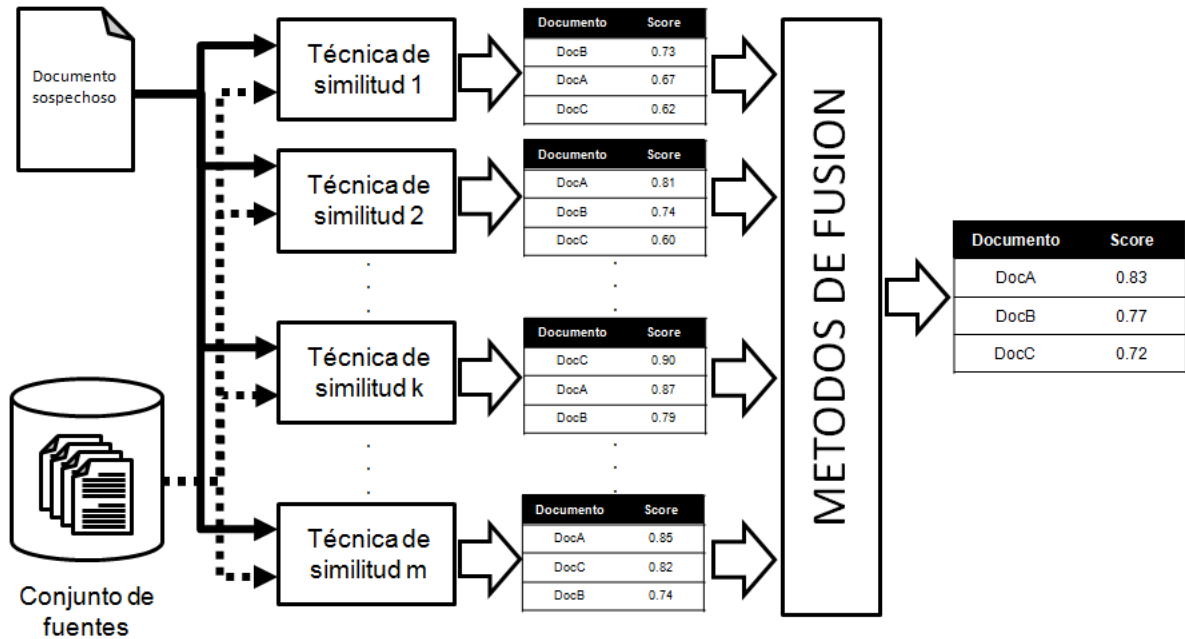


Figura 2.6: Sistemas de fusión de datos

En la figura 2.7 se explica lo mencionado; se tienen dos casos de sistemas de detección de plagio (a) y (b), en ambos casos se tiene un universo con todos los documentos plagiados y tres sistemas de detección de plagio (A, B y C). En el **caso (a)** se presentan los sistemas de detección individuales: El *Sistema A* y el *Sistema B* pueden reconocer documentos en común ( $A \cap B$ , intersección de los *Sistemas A* y *B*) y otros tantos que sólo fueron reconocidos por el *Sistema A* ( $A - B$ ) o por el *Sistema B* ( $B - A$ ); mientras que el *Sistema C* reconoce documentos que ni el *Sistema A*, ni el *Sistema B* ( $(A \cup B) \cap C = \emptyset$ ). En el **caso (b)** se presentan los sistemas de detección de plagio fusionados: La detección de cada sistema por separado sigue siendo la misma pero al tenerlos unidos ( $A \cup B \cup C$ ) el rango aumenta, de este modo se consigue una mejor detección de documentos plagiados para el mismo universo, describiendo mejor este caso, los *Sistemas* dejan de ser nombrados como A, B o C y llegan a ser un nuevo sistema fusión (F) tal que  $F = A \cup B \cup C$ .

Haciendo una revisión bibliográfica, en la literatura se encuentran trabajos de autores que ya han ahondado en algunos métodos de fusión de punjates, que genéricamente se denominan *métodos de fusión de datos*. A continuación se presentarán los beneficios de desarrollar un método de fusión de datos y posteriormente se describirán algunos *métodos para fusión de datos*, los cuales están divididos en dos tipos:

- Métodos de fusión de datos basados en ranking
- Métodos de fusión de datos basados en score

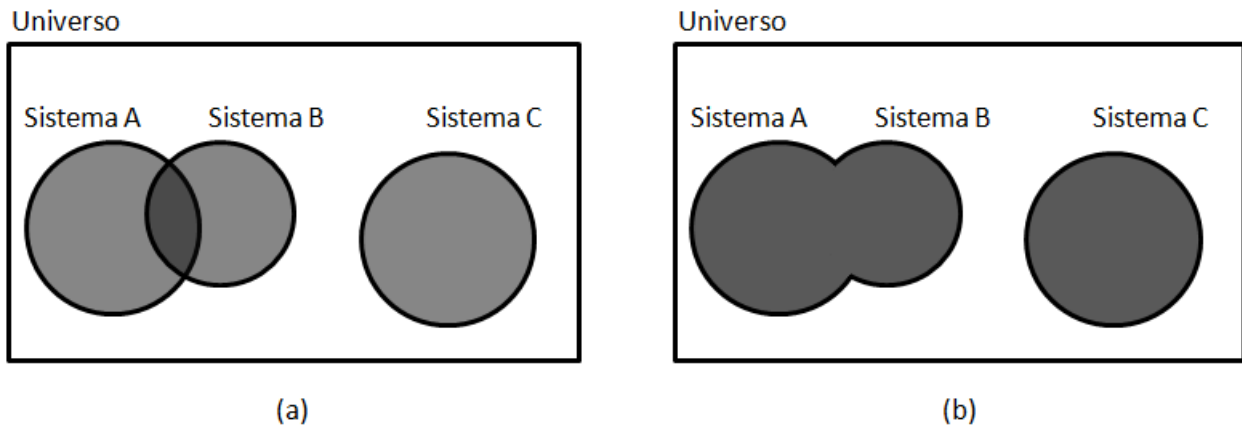


Figura 2.7: (a) Sistemas de detección individualmente. (b) Sistemas de detección fusionados

#### 2.4.1. Beneficios de un método de fusión de datos

Hasta este punto hemos discutido la ventaja general de un sistema de fusión de datos, como un sistema que permite combinar todas las fortalezas de los diversos sistemas de detección de similitud. En esta sub-sección, así como lo plantea Mark Montague en sus tesis [48] y en el trabajo conjunto con Javed Aslam [49] es necesario explicar los beneficios de contar con un método de fusión de datos.

##### 1. Mejora en la recuperación (RECALL)

La recuperación (RECALL) es el valor que indica el número de documentos relevantes que reconoció el sistema efectivamente, en nuestro caso será: *De todos los documentos plagiados existentes en la base de datos, cuantos de ellos reconoció el sistema.* En otras palabras, un **alto RECALL** indicará que el sistema reconoció a muchos de los documentos plagiados que existen en la base; análogamente un **bajo RECALL** indica que, el sistema reconoció muy pocos documentos plagiados del universo de prueba. De acuerdo al trabajo de Joon Lee [38, 39], un sistema de fusión de datos logra un alto RECALL porque los diversos sistemas de entrada que trabajan sobre la misma base de datos entregan no sólo distintos documentos relevantes sino que le agrega distintas relevancias (por score o ranking) que en conjunto ayuda al sistema de fusión de datos a reconocer los documento más relevantes, dejando de lado aquellos documentos irrelevantes.

##### 2. Mejora en la precisión (PRECISION)

La precisión (PRECISION) indica, del grupo de documentos que el sistema encontró como relevantes, el número que es realmente relevante. Para el caso del plagio indicará: *De todos los documentos que el sistema dijo que son plagiados cuantos son efectivamente plagiados.* Es decir, una **alta PRECISION** indica que de todos los documentos donde el sistema detectó plagio muchos de ellos si eran plagiados; análogamente una **baja PRECISION** indica que el sistema reconoció documentos pero muy pocos eran efectivamente plagiados. Ng y Kantor en [51] concluyen que un sistema de fusión de datos puede mejorar su precisión por sus datos de entrada. Si un documento es reconocido como relevante por



muchos sistemas de entrada esto indicará que efectivamente será relevante en la salida.

### 3. **Consistencia en el desempeño (PERFORMANCE)**

En la investigación realizada por Erik Selberg y Oren Etzioni [61], ellos muestran que para los MetaSearches<sup>23</sup> una misma consulta (query) en diversos motores de búsqueda puede tener resultados distintos (considerando que la base de datos, en este caso la Web, es la misma para todos los motores de búsqueda) siendo el desempeño de algunos buscadores bueno y en otros muy pobre. La misma lógica ocurre con los métodos de similitud de documentos; para un mismo par de documentos (Sospechoso – Fuente k-ésima) un método puede tener un buen desempeño, mientras que otro no, lo cual ocasiona inconsistencia en el funcionamiento. En cambio, un sistema de fusión de datos siempre funcionará con un desempeño promedio (ni muy eficiente, ni poco eficiente) logrando así una mejor consistencia.

### 4. **Arquitectura por módulos (MODULAR)**

Un sistema de fusión de datos desde su diseño toma en consideración que se le puedan agregar o quitar métodos de entrada, conforme se quiera agilizar o agregar especificidad al proceso respectivamente. Cada método de entrada se define como un módulo. Es debido a esta característica de tener módulos, la cual no posee ningún método de similitud de copia, que el sistema de fusión de datos se torna potente gracias a su flexibilidad y comodidad de adecuarlo al cambio.

---

<sup>23</sup>Esta es una aplicación específica de los métodos de fusión. Los MetaSearches son buscadores que no tienen base de datos interna, se encargan de buscar una consulta en otros buscadores y mediante un *sistema de fusión de datos* unir los resultados para la *query* ingresada por el usuario, la fusión entrega un único resultado que indica las páginas más relevantes a la consulta.

## 2.4.2. Métodos de fusión de datos basados en ranking

El objetivo de la *fusión basada en ranking* es asignar un score a cada elemento de los *vectores respuesta* que se originan por los *métodos de detección de plagio* que ingresan al sistema de fusión. Y que posteriormente integrándolos, sin pérdida de generalización, se tenga un nuevo *vector respuesta* el cual al ser ordenado descendentemente sea nuestro resultado del RANKING-FUSION. [42, 79]

En este contexto, los *vectores respuesta* de cada método de detección de plagio se denominarán *listas ordenadas*, además se utilizará la siguiente notación:

- $Spc$  : Documento sospechoso analizado
- $Src$  : Universo de documentos fuentes para un documento sospechoso.  $|Src| = n, \quad n \in \mathbb{Z}^+$
- $Mtd$  : Universo de sistemas de similitud de documentos.  $|Mtd| = m, \quad m \in \mathbb{Z}^+$
- $k$  : k-ésimo método de detección de similitud de documentos.  $k = 1, 2, 3, \dots, m$
- $S_k$  : Subconjunto de documentos fuentes con mayor similitud a  $Spc$  que entregó el k-ésimo método de detección de similitud de documentos.  $S_k \subseteq Src$
- $x_i^k$  : Elemento i-ésimo del subconjunto  $S_k$ .  $x_i^k \in S_k$
- $\tau^k$  : Lista ordenada (rankeada) de elementos  $x_i^k$  que pertenecen a  $S_k$ .  
 $\tau^k = [x_{i_1}^k \succeq x_{i_2}^k \succeq \dots \succeq x_{i_j}^k \succeq \dots \succeq x_{i_{p-1}}^k \succeq x_{i_p}^k], \quad p \leq n$
- $\tau^F$  : Lista ordenada que contiene la fusión de las k-listas ordenadas  $\tau^k$ .  
 $\tau^F = \{x_i^k | Fusion(x_i^k), x_i^k \in \tau^k, \forall k \in Mtd\}$

En la penúltima expresión, el símbolo  $\succeq$  indica alguna relación de ordenamiento para los elementos de  $S_k$ . Siendo preferido el elemento  $x_i^k$  que se encuentre a la izquierda de  $\succeq$  sobre el elemento que se encuentre a la derecha. Lo anterior permite definir que el elemento  $x_i^k$  presentado en  $\tau$  indica la posición, o ranking, de  $x_i^k$ . En otras palabras, el elemento  $x_{i_1}^k$  tiene más probabilidad de copia que el elemento  $x_{i_p}^k$ . Además, aclarar que la expresión “ $\tau^k(x_i^k)$ ” indica el ranking del elemento  $x_i^k$  en la lista  $\tau$  entregada por el método de  $k$ .

El ordenamiento basado en ranking puede expresarse como una función objetivo porque se busca un sistema de fusión de datos que entregue una lista ordenada  $\tau^F$  que minimice las diferencias con los  $\tau^k$  de entrada.

Formalizando las expresiones, de acuerdo al trabajo de Yu-Ting Liu [42], se puede representar la fusión de datos como la lista  $\tau^F$  y la inequación

$$H \cdot \tau^F < 0 \tag{2.1}$$

De acuerdo a Liu,  $H$  es una matriz que representa la modificación de pesos para hallar una buena relación entre elementos. Del ejemplo de Liu, sea:  $\tau^F = (x_1, x_2, x_3, x_4)$  y de las listas  $\tau^k$  de entrada indican

que el elemento 1 debe ser rankeado mejor que el elemento 2, además que el elemento 4 debe ser rankeado mejor que el elemento 3. Entonces la inecuación se expresará:

$$H \cdot \tau^F < 0$$

Donde:

$$H = \begin{bmatrix} -1 & 1 & 0 & 0 \\ 0 & 0 & 1 & -1 \end{bmatrix} \quad (2.2)$$

La expresión 2.1 no garantiza que exista una solución que satisfaga todas las listas  $\tau^k$  ingresadas al sistema, sin embargo se puede resolver introduciendo una variable de holgura  $t$  que represente el error:

$$H \cdot \tau^F < t, \quad t \geq 0 \quad (2.3)$$

Entonces, para reducir el error en el resultado rankeado basta con minimizar la norma de  $t$ , lo cual puede ser expresado formalmente como el siguiente problema de optimización:

$$\begin{aligned} & \min_{(\tau^F, \alpha, t)} \quad t^\top \cdot t \\ \text{s.a.} \quad & \tau^F = \text{Fusion} \{ \tau^1, \dots, \tau^k, \dots, \tau^m, \alpha \} \\ & \alpha \in S_k \quad \forall k \in Mtd \\ & H \cdot \tau^F < t \\ & t \geq 0 \end{aligned} \quad (2.4)$$

Donde  $\text{Fusion} \{ \tau^1, \dots, \tau^k, \dots, \tau^m, \alpha \}$  es el método de fusión que une las listas  $\tau^k$  y que utiliza un ponderador interno  $\alpha$ , siendo  $\alpha$  un vector de parámetros para la región factible del universo de documentos recuperados ( $S_k$ ). Además, la función objetivo  $t^\top \cdot t$  indica la no similitud entre el conjunto fusión  $\tau^F$  y los métodos ingresados  $\tau^k$ . Donde el valor 0, de la función objetivo, indica que  $\text{Fusion} \{ \tau^1, \dots, \tau^k, \dots, \tau^m, \alpha \}$  satisface todos los  $\tau^k$  y las restricciones de 2.4.

Existen distintos modos para resolver y optimizar el sistema de fusión, las cuales se presentarán a continuación divididas en: *técnicas basadas en distancia*, en *técnicas basadas en lógica difusa* y en *técnicas algorítmicas*.

#### ■ Técnicas basadas en similitud

Las técnicas basadas en similitud, como su nombre lo indica, se encargan de representar numéricamente la similitud entre listas rankeadas; en otras palabras, entregan un valor ordinal que indica la correlación entre un par de listas rankeadas. Estas técnicas de similitud por si mismas no pueden determinar una lista fusionada  $\tau^F$ , sin embargo sirven como apoyo estadístico para que el usuario pueda conocer cuales listas ordenadas  $\tau^k$  son más similares entre si y con ello definir una *lista fusionada*

aproximada  $\tau^{F^*}$  que cumpla con los criterios de la expresión 2.4. Por ejemplo, si se tienen las listas  $\tau^1, \tau^2, \tau^3, \tau^4, \tau^5$  y mediante alguna técnica de similitud se encuentra que las listas  $\tau^1, \tau^3, \tau^4, \tau^5$  son muy similares (o iguales) entre sí, entonces la *lista fusionada aproximada*  $\tau^{F^*}$  a definir tendrá que ser como ellas y de ese modo asegurar que resume de buena manera todas las listas  $\tau^i$ , ( $i = 1, 2, 3, 4, 5$ ). Por otro lado, se supone que la lista  $\tau^2$  es un ranqueo no óptimo, por ello no presenta una alta *similitud* con las otras listas y consecuentemente tampoco debe ser similar a  $\tau^{F^*}$ .

- **Spearman footrule distance.** Este indicador de distancia resulta de calcular la suma de las distancias absolutas entre los rankings de un mismo elemento  $x_i$  de acuerdo a dos listas  $\tau^k$  y  $\tau^l$  [23]. Formalmente, la distancia Spearman footrule  $D_{Foot}(\tau^k, \tau^l)$  será:

$$D_{Foot}(\tau^k, \tau^l) = \sum_{i=1}^{|S_k|} \left| \tau^k(x_i^k) - \tau^l(x_i^l) \right| \quad (2.5)$$

Además:

$$\tau^k(x_i^k) = 0 \leftrightarrow x_i^k \notin S_k, \quad \forall k = 1, 2, \dots, m.$$

Si se divide por el máximo valor que es  $\lfloor |S_k|^2/2 \rfloor$  se obtiene un valor normalizado para la *Spearman footrule distance* que va entre 0 y 1 [25], [8].

Donde:

Si:  $D_{Foot}(\tau^k, \tau^l)^{normalizado} \rightarrow 0 \Rightarrow$  Las listas son totalmente idénticas.

Si:  $D_{Foot}(\tau^k, \tau^l)^{normalizado} \rightarrow 1 \Rightarrow$  Las listas son totalmente diferentes.

La ecuación 2.5 sólo funciona para dos listas, sin embargo se puede tener una expresión generalizada del siguiente modo: Sea  $p$  el número de listas parciales  $\tau^k$  ( $k = 1, 2, \dots, p-1, p$ ;  $p \in \mathbb{Z}^+$ ) y sea  $\tau$  una lista ordenada con respecto al universo  $Src$ . Entonces la distancia  $D_{Foot}(\tau, \tau^1, \tau^2, \dots, \tau^{p-1}, \tau^p)$  cumplirá:

$$D_{Foot}(\tau, \tau^1, \tau^2, \dots, \tau^{p-1}, \tau^p) = \sum_{k=1}^p D_{Foot}(\tau, \tau^k) / p \quad (2.6)$$

- **Kendall tau distance.** Este es un medidor de distancia entre dos listas  $\tau^k$  y  $\tau^l$ ; también es conocido como *distancias del ordenamiento de burbuja*<sup>24</sup> porque calcula el número de pasos que debe ejecutar el *algoritmo de burbuja* para que una lista se parezca a la otra [35].

Sean dos listas  $\tau^k$  y  $\tau^l$ , la *distancia de Kendall tau*  $D_{Ken}(\tau^k, \tau^l)$  estará dada por todos los pares ordenados  $(x_i, x_j)$  que tienen una relación de precesión que cumple que el elemento  $x_i$  estará en una peor posición que el elemento  $x_j$  en la lista  $k$ , y a la vez tiene una mejor posición que el mismo  $x_j$  en la lista  $l$  [25]:

$$D_{Ken}(\tau^k, \tau^l) = \sum_{1 \leq i \leq j \leq |S_k|} \left| \left\{ (x_i, x_j) \mid x_i < x_j, (x_i^k \prec x_j^k) \wedge (x_i^l \succ x_j^l) \right\} \right| \quad (2.7)$$

<sup>24</sup>BubbleSort distance, por su nombre en inglés. El *BubbleSort* es un algoritmo de ordenamiento que, a partir de una lista desordenada de elementos con pesos, re-ubica los elementos con valores mas pesados al final de la lista, mientras los más livianos los coloca al inicio.

Además:

$$\tau^k(x_i^k) = 0 \leftrightarrow x_i^k \notin S_k, \quad \forall k = 1, 2, \dots, m.$$

Si el valor de distancia  $D_{Ken}(\tau^k, \tau^l)$  se divide por el número  $\binom{|S_k|}{2}$  se obtiene un valor normalizado para la *Kendall tau distance* que va entre 0 y 1 [8, 25].

Donde:

Si:  $D_{Ken}(\tau^k, \tau^l)^{normalizado} \rightarrow 0 \Rightarrow$  Las listas  $\tau^k$  y  $\tau^j$  tienen el mismo ordenamiento.

Si:  $D_{Ken}(\tau^k, \tau^l)^{normalizado} \rightarrow 1 \Rightarrow$  Las listas  $\tau^k$  y  $\tau^j$  tienen un ordenamiento opuesto.

Similar al *Spearman footrule distance* (Ver ecuación 2.6), se tiene una generalización para el *Kendall tau distance*:

$$D_{Ken}(\tau, \tau^1, \tau^2, \dots, \tau^{p-1}, \tau^p) = \sum_{k=1}^p D_{Ken}(\tau, \tau^k)/p \quad (2.8)$$

#### ■ Técnicas basadas en lógica difusa

Estas técnicas exploran el uso de metodologías de lógica difusa para encontrar el mejor ordenamiento para un conjunto de listas de ingreso  $\tau^k$ .

- **Ordenamiento difuso por la técnica de Shimura.** Esta es una de las técnicas de ordenamiento difuso más antigua en la bibliografía, data del año 1973 y fue propuesta por Masamichi Shimura [64]. Para ésta técnica se hace uso de la definición de la *función de membresía* ( $f_{x_j}(x_i)$ ) la cual indica la preferencia de un elemento  $x_i$  sobre un elemento  $x_j$  (Donde:  $x_i, x_j \in Src$ ).

$$f_{x_j}(x_i) = \frac{\left| \left\{ x_i^k, \forall k \in Mtd \mid x_i^k \succ x_j^k, \forall x_j^k \in S_k, i \neq j \right\} \right|}{|Mtd|} \quad (2.9)$$

De otro modo, la ecuación 2.9 puede expresarse como:

$$f_{x_j}(x_i) = \frac{\sum_{k=1}^{|Mtd|} \mathbf{1}_{(x_i^k \succ x_j^k)}}{|Mtd|} \quad (2.10)$$

Donde:

$$f_{x_j}(x_i) = 1, \quad \forall i = j$$

Luego, ordenando descendientemente las *funciones de membresía* encontradas, el ranking para el  $i$ -ésimo elemento ( $x_i$ ) [43] será:

$$C_i = \min_{j=1}^{|Src|} f(x_i|x_j) \quad (2.11)$$

Donde la expresión  $f(x_i|x_j)$  es una matriz de  $|Src| \times |Src|$ , con el  $i$  en las filas y  $j$  en las columnas que cumple:

$$f(x_i|x_j) = \frac{f_{x_j}(x_i)}{\max(f_{x_j}(x_i), f_{x_i}(x_j))} \quad (2.12)$$

La lista fusionada  $\tau^F$  utilizará los valores  $C_i$  ordenados del siguiente modo:

$$\tau^F = \{x_i | C_i \leq C_j; i, j = 1, 2, \dots, |Src| - 1, |Src|\} \quad (2.13)$$

Además, para que  $\tau^F$  tenga un buen ordenamiento es bueno recalcar que  $C_i$  está ordenado descendientemente y por lo tanto cumple lo siguiente:

$$C_1 \leq C_2 \leq \dots \leq C_{(i-1)} \leq C_i \leq C_{(i+1)} \leq \dots \leq C_{(|Src|-1)} \leq C_{|Src|}$$

- **Ordenamiento difuso por la técnica de Dubois y Prade (DP).** Esta técnica está basada en la distribución Gaussiana, es decir, considera la posición media y su varianza de los elementos  $x_i^k$  respecto a una lista  $k$  [24, 56]. Siendo la media y la varianza del elemento  $x_i$  expresada como  $\bar{X}_{x_i}$  y  $\bar{\sigma}_{x_i^2}$  respectivamente. Donde:

$$\bar{X}(x_i) = \frac{\sum_{k=1}^{|Mtd|} \tau^k(x_i^k)}{|Mtd|} \quad (2.14)$$

$$\bar{\sigma}^2(x_i) = \frac{\sum_{k=1}^{|Mtd|} (\tau^k(x_i^k) - \bar{X}(x_i))^2}{|Mtd|} \quad (2.15)$$

Utilizando las ecuaciones 2.14 y 2.15, se construye la expresión de ordenamiento difuso DP la cual sus autores llamaron *función de membresía Gaussiana*:

$$\mu(x_i, p) = \frac{1}{\sqrt{2\pi\bar{\sigma}^2(x_i)}} e^{\left[-\frac{1}{2} \left(\frac{(p-\bar{X}(x_i))^2}{\bar{\sigma}^2(x_i)}\right)\right]} \quad (2.16)$$

Donde:

$$\begin{aligned} p &= 1, 2, \dots, |S| \\ |S| &= |S_k|, \quad \forall k \in Mtd \\ S &= S_k = S_l, \quad \forall k, l \in Mtd \end{aligned}$$

La expresión de DP (Ecuación 2.16) considera que los sistemas de entrada trabajan para un mismo universo de elementos  $S \subseteq Src$  y sólo se encargan de darles un ordenamiento particular. Por ejemplo, sea el universo  $S = a, b, c, d, e$  entonces las listas  $\tau^k$  sólo ordenarán a partir de ellos como  $\tau^1 = e, b, d, a, c$ ,  $\tau^2 = c, b, d, a, e$ ,  $\tau^3 = b, c, a, d, e$ ; nótese que  $S_1 = S_2 = S_3 = S$  y consecuentemente  $|S_1| = |S_2| = |S_3| = |S|$ . Volviendo a la ecuación 2.16,  $p$  indica la posición que puede tomar el elemento  $x_i$  dentro de alguna lista  $\tau$ .

Luego, para conocer que verdaderamente el elemento  $x_i$  está en una mejor posición que  $x_j$  ( $x_i \succeq x_j$ ) se utiliza:

$$DP(x_i \succeq x_j) = \max_{(\forall p \geq q)} \{ \min [\mu(x_i, p), \mu(x_j, q)] \}, \quad \forall i \neq j \quad (2.17)$$

Entonces, generalizando la expresión 2.17, para que el elemento  $x_i$  esté en la mejor posición versus todos los otros elemento  $x_j$  ( $\forall x_j \in S$ ) debe cumplir la expresión 2.18:

$$DP(x_i \succeq \{\forall x_j \in S\}) = \min\{DP(x_i \succeq x_1), DP(x_i \succeq x_2), \dots, DP(x_i \succeq x_{(i-1)}), DP(x_i \succeq x_{(i+1)}), \dots, DP(x_i \succeq x_{(|S|-1)}), DP(x_i \succeq x_{|S|})\} \quad (2.18)$$

Y finalmente, la lista fusionada será elegida a partir del ordenamiento ascendente para los resultados de  $DP(x_i \succeq \{\forall x_j \in S\})$ , del siguiente modo:

$$\tau^F = \{x_i | DP(x_i \succeq \{\forall x_r \in S\}) \leq DP(x_j \succeq \{\forall x_r \in S\}), i \neq j, \{x_i, x_j\} \in S\} \quad (2.19)$$

Donde:

$$DP(x_{i_1} \succeq \{\forall x_j \in S\}) \leq DP(x_{i_2} \succeq \{\forall x_j \in S\}) \leq \dots \leq DP(x_{i_{(|S|-1)}} \succeq \{\forall x_j \in S\}) \leq DP(x_{i_{|S|}} \succeq \{\forall x_j \in S\})$$

- **Técnica mean-by-variance (MBV).** Esta técnica se apoya en los cálculos de la media (Ecuación 2.14) y la varianza (Ecuación 2.15).

$$mbv(x_i) = \left( \frac{\bar{X}(x_i)}{\bar{\sigma}^2(x_i)} \right) \quad (2.20)$$

Luego, el ordenamiento ascendente para obtener el ranking fusionado ( $\tau^F$ ):

$$\tau^F = \{x_i | mbv(x_i) \leq mbv(x_j); \{x_i, x_j\} \in |S|\} \quad (2.21)$$

Donde se cumple:

$$mbv(x_1) \leq \dots \leq mbv(x_{(i-1)}) \leq mbv(x_i) \leq mbv(x_{(i+1)}) \leq \dots \leq mbv(x_{|Src|})$$

$$\begin{aligned} |S| &= |S_k|, & \forall k \in Mtd \\ S &= S_k = S_l, & \forall k, l \in Mtd \end{aligned}$$

- **Técnica de Shimura modificada.** Esta técnica es propuesta en el trabajo de Sufyan Beg y Nesar Ahmad [8], la cual consiste en cambiar el valor de  $C_i$ , presentado en la ecuación 2.11, por una nueva expresión  $C'_i$  que incluyera pesos difusos  $w_i$ :

$$C'_i = \sum_{j=1}^{|Src|} w_j z_j \quad (2.22)$$

Donde:  $w_j$  : Peso difuso del elemento  $x_j$   
 $z_j$  : valor del j-ésimo elemento en la fila i-ésima de la matriz  $f(x_i | x_j)$

Para hallar el valor de  $z_i$  se sigue la siguiente expresión:

$$z_i = \max_{j=1}^{|Src|} f(x_i | x_j) \quad (2.23)$$

Para definir el valor de  $w_i$  se recurre al concepto de **Ordered Weighted Averaging (OWA)** presentado por Ronald Yager en 1988 [80], con lo cual se hace uso de un cuantificador difuso  $Q$ .

$$w_i = Q\left(\frac{i}{m}\right) - Q\left(\frac{(i-1)}{m}\right), i = 1, 2, \dots, |Src| \quad (2.24)$$

$$Q(0) = 0$$

Donde,  $Q(x)$  es una función que depende del valor de  $x$  y además de unas cotas de posición  $a$  y  $b$  ( $\forall a, b \in [0; 1]$ ) además  $x, a, b \in [0; 1]$ . De acuerdo a Beg y Ahmad [8] el rango de  $[a; b]$  puede ser de tres tipos (Ver Cuadro 2.1):

$$Q(x) = \begin{cases} 0, & \text{si } x < a \\ \frac{x-a}{b-a} & \text{si } a \leq x \leq b; \\ 1, & \text{si } x > b \end{cases} \quad \forall x, a, b \in [0; 1] \quad (2.25)$$

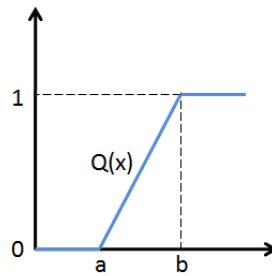


Figura 2.8: Función para el cuantificador difuso Q

[a;b]	COTAS
[0,3;0.8]	<i>Algunos</i>
[0;0.5]	<i>Al menos la mitad</i>
[0.5;1]	<i>Tantos como sean posibles</i>

Cuadro 2.1: Cotas para el cuantificador difuso Q

### ■ Técnicas algorítmicas

Esta clase de métodos de fusión, son las que utilizan una lógica matemática propia, usualmente algorítmica, para encontrar la mejor lista fusionada  $\tau^F$  a partir de las listas de entrada  $\tau^k$ .

- **Técnica de minimización entrópica.** La minimización entrópica se aplica para sistemas donde la cantidad de datos es grande y estática [56], en este caso se requiere que se tenga una gran cantidad de documentos fuentes ( $Src$ ) y métodos de detección de plagio ( $Mtd$ ). Para esta técnica se asume que la posición ( $\tau(x_i) = p$ ) de un conjunto de documentos ( $x_i \in S; S = S_k, \forall k \in Mtd$ ) se encuentra dentro del rango  $[p_1, p_2]$ ; entonces se podría separar dos sub-conjuntos  $\alpha$  y  $\beta$  respecto a  $x_i$  que cumplan:  $\alpha_{x_i} = [p_1, p]$  y  $\beta_{x_i} = [p, p_2]$ .

Conforme se mueva la posición  $\tau(x_i) = p$  entre  $p_1$  y  $p_2$  se irá calculando el valor de su entropía. Aquel valor de  $p$  que entregue la mínima entropía será la posición más apropiada del documento  $x_i$ .

El cálculo entrópico para un documento  $x_i$  cuya posición  $p$  se encuentra entre  $p_1$  y  $p_2$  será:



$$E_{x_i}(p) = p_{x_i}(p)Sp_{x_i}(p) + q_{x_i}(p)Sq_{x_i}(p) \quad (2.26)$$

Donde:

$$Sp_{x_i}(p) = - [pm_{x_i}(p) \ln (pm_{x_i}(p)) + pr_{x_i}(p) \ln (pr_{x_i}(p))] \quad (2.27)$$

$$Sq_{x_i}(p) = - [qm_{x_i}(p) \ln (qm_{x_i}(p)) + qr_{x_i}(p) \ln (qr_{x_i}(p))] \quad (2.28)$$

Además, Ross [56] hace un listado de otras ecuaciones que resultan necesarias para un cálculo correcto:

$$pm_{x_i}(p) = \frac{rm_{x_i}(p) + 1}{rm_{x_i}(p) + rr_{x_i}(p) + 1} \quad (2.29)$$

$$pr_{x_i}(p) = \frac{rr_{x_i}(p) + 1}{rm_{x_i}(p) + rr_{x_i}(p) + 1} \quad (2.30)$$

$$qm_{x_i}(p) = \frac{Rm_{x_i}(p) + 1}{Rm_{x_i}(p) + Rr_{x_i}(p) + 1} \quad (2.31)$$

$$qr_{x_i}(p) = \frac{Rr_{x_i}(p) + 1}{Rm_{x_i}(p) + Rr_{x_i}(p) + 1} \quad (2.32)$$

$$p_{x_i}(p) = \frac{rm_{x_i}(p) + rr_{x_i}(p)n}{r} \quad (2.33)$$

$$q_{x_i}(p) = 1 - p_{x_i}(p) \quad (2.34)$$

$$Rm_{x_i}(p) = \left| \left\{ x_i^k \mid \tau^k(x_i^k) \in [p_1 + p; p_2]; x_i \in S_k; \forall k \in Mtd \right\} \right| \quad (2.35)$$

$$Rr_{x_i}(p) = \left| \left\{ x_j^k \mid \tau^k(x_j^k) \in [p_1 + p; p_2]; x_j \in S_k; j \neq i; \forall k \in Mtd \right\} \right| \quad (2.36)$$

$$rm_{x_i}(p) = \left| \left\{ x_i^k \mid \tau^k(x_i^k) \in [p_1; p_2 + p]; x_i \in S_k; \forall k \in Mtd \right\} \right| \quad (2.37)$$

$$rr_{x_i}(p) = \left| \left\{ x_j^k \mid \tau^k(x_j^k) \in [p_1; p_2 + p]; x_j \in S_k; j \neq i; \forall k \in Mtd \right\} \right| \quad (2.38)$$

$$r = |\{x_j \mid x_j \in S; S = S_k; \forall k \in Mtd\}| \quad (2.39)$$

La lista fusionada final  $\tau^F$  será un ordenamiento de elementos  $x_i$  que en la posición  $p$  ascendente que entregue el mínimo valor de entropía  $E_{x_i}(p)$

$$\tau^F = \{x_i \mid E_{x_i}(p) \leq E_{x_j}(p); p = 1, 2, \dots, |Src|; \{x_i, x_j\} \in S\} \quad (2.40)$$

- **Borda count.** Este método asume como base que ninguna lista  $\tau^k$  es igual de *importante* con otra lista  $\tau^l$  (Donde:  $k \neq l, \forall k, l \in Mtd$ ) [72]. Utilizando la expresión 2.4 (Página 28), se tendrá que los diversos métodos de detección de copia ingresados poseerán pesos; a partir de esto último se define la ecuación de fusión de Borda:

$$\tau^F = Fusion \left\{ \tau^1, \dots, \tau^k, \dots, \tau^m, \alpha \right\} = \sum_{k=1}^m \alpha_k \cdot \tau^k \quad (2.41)$$

Con lo cual, se obtiene un problema de optimización de la forma:

$$\begin{aligned} & \min_{(\tau^F, \alpha, t)} t^\top \cdot t \\ \text{s.a.} \quad & \tau^F = \sum_{k=1}^m \alpha_k \cdot \tau^k \\ & \sum_{k=1}^m \alpha_k = 1 \\ & \alpha_k \geq 0 \quad \forall k \in Mtd \\ & H \cdot \tau^F < t \\ & t \geq 0 \end{aligned} \quad (2.42)$$

Cabe resaltar que en este caso particular, el vector de parámetros  $\alpha_k$  indica los pesos para cada lista rankeada  $\tau^k$ .

Al resolver la expresión 2.42, se llegará a una nueva ecuación que llamada el Borda Score  $B(x_i)$  [81]. El Borda Score resuelve encontrar un puntaje para todo elemento  $x_i \in Src$ .

$$B(x_i) = \left| \left\{ x_i^k, \quad \forall k \in Mtd \mid (x_i^k \succ x_j^k), x_i^k \in S_k, \forall x_j^k \in S_k, i \neq j \right\} \right| \quad (2.43)$$

Donde:

$$B(x_i) = 0, \quad \forall x_i \notin \bigcup_{k=1}^m \{S_k\}$$

Finalmente, ordenando descendentemente los resultados de  $B(x_i)$  para todo  $x_i \in Src$  se tendrá la lista fusionada ( $\tau^F$ ) buscada.

$$\tau^F = \{x_i \mid B(x_i) \leq B(x_j); \{x_i, x_j\} \in |S|\} \quad (2.44)$$

Donde se cumple:

$$\begin{aligned} B(x_1) &\leq \dots \leq B(x_{(i-1)}) \leq B(x_i) \leq B(x_{(i+1)}) \leq \dots \leq B(x_{|Src|}) \\ |S| &= |S_k|, \quad \forall k \in Mtd \\ S &= S_k = S_l, \quad \forall k, l \in Mtd \end{aligned}$$

### 2.4.3. Métodos de fusión de datos basados en score

Los métodos de fusión de datos basados en score funcionan en dos etapas: *Linealización* y *combinación* [74].

La *etapa de linealización* se encarga de tomar todos los *scores* originados por los  $k$  diversos sistemas de clasificación ( $score \in [a, b]$ ;  $a, b \in \mathbb{R}$ ), tal como se explicó en la Sección 2.4 de la página 22, y lo acota en un rango que usualmente va entre 0 a 1 ( $score \in [0, 1]$ ). Las técnicas de linealización más utilizadas hoy en día fueron presentadas por primera vez en el trabajo de Mark Montague y Javed Aslam [49].

La *etapa de combinación* viene inmediatamente después de la *etapa de linealización*, esta etapa se encarga propiamente de fusionar los resultados de las diversas listas en una única lista. Para realizar la fusión, o combinación como también es llamada, existen diversos métodos que fueron propuestos por Joseph Shaw y Edward Fox en 1993 [63].

Para las siguientes técnicas se utilizará la misma notación presentada para el ranking en la sección 2.4.2 (página 27) pero se considerará adicionalmente:

- $\rho^k$  : Lista que contiene los elementos  $x_i^k$  ordenados ascendentemente por su score.  
 $\rho^k = [x_{i_1}^k \succeq x_{i_2}^k \succeq \dots \succeq x_{i_j}^k \succeq \dots \succeq x_{i_{p-1}}^k \succeq x_{i_p}^k]$ ,  $p \leq n$
- $\rho^F$  : Lista ordenada que contiene la fusión de las  $k$ -listas ordenadas  $\rho^k$ .  
 $\tau^F = \{x_i^k | Fusion(x_i^k), x_i^k \in \rho^k, \forall k \in Mtd\}$
- $Sc_k(x_i)$  : Score del elemento  $x_i$  entregado por el método  $k$ . Es equivalente a decir:  $Sc_k(x_i) = \rho^k(x_i^k)$ , para tener una notación menos densa se prefiere el  $Sc_k(x_i)$ .
- $Sc_F(x_i)$  : Score fusionado para el elemento  $x_i$ . Este puntaje es obtenido luego de haber utilizado una técnica de fusión de scores.

#### ■ Técnicas de linealización

- **Estándar.** Esta es la técnica más simple de linealización, se encarga sólo de cambiar la escala de los scores de  $[a; b] \in \mathbb{R}$  a una escala exacta de  $[0; 1]$ .

$$\overline{Sc}_k(x_i) = \frac{Sc_k(x_i) - ScMin_k}{ScMax_k - ScMin_k}; \quad x_i \in S_k; \forall k \in Mtd \quad (2.45)$$

Donde:

$$ScMin_k = \{Sc_k(x_i) | Sc_k(x_i) \leq Sc_k(x_j); \forall x_i, x_j \in S_k\} \quad (2.46)$$

$$ScMax_k = \{Sc_k(x_i) | Sc_k(x_i) \geq Sc_k(x_j); \forall x_i, x_j \in S_k\} \quad (2.47)$$

- **SUM.** La técnica de *linealización de suma* es similar a la *linealización estándar* (Ecuación 2.45) con la diferencia que en no utiliza el valor del *score máximo* ( $ScMax$ ) como en la definición

estándar sino que utiliza un score máximo que es la suma de todos los scores obtenidos por el método k-ésimo.

$$\overline{Sc}_k(x_i) = \frac{Sc_k(x_i) - ScMin_k}{ScSum_k - ScMin_k}; \quad x_i \in S_k; \forall k \in Mtd \quad (2.48)$$

Donde:

$$ScSum_k = \sum_{\forall x_i \in S_k} (Sc_k(x_i)) \quad (2.49)$$

El valor de  $ScMin_k$  utilizado en la ecuación 2.48 es el mismo de la ecuación 2.46.

- **ZMUV.** También conocida como **Z-Score**, es una técnica que utiliza valores de la media ( $\mu_k$ ) y la varianza ( $\sigma_k$ ) de scores que entrega el k-ésimo método de detección de plagio.

$$\overline{Sc}_k(x_i) = \frac{Sc_k(x_i) - \mu_k}{\sigma_k}; \quad x_i \in S_k; \forall k \in Mtd \quad (2.50)$$

Donde:

$$\mu_k = \frac{ScSum_k}{|Mtd|} \quad (2.51)$$

$$\sigma_k = \frac{\sum_{\forall x_i \in S_k} (Sc_k(x_i) - \mu_k)^2}{|Mtd|} \quad (2.52)$$

- **Técnica de la teoría de la información de Shannon.** Esta técnica fue planteada por Yu Suzuki, et. al. [67] en el años 2005, en la cual propone tres pasos de normalización basados en la teoría de la información de Shannon, particularmente conceptos de entropía y valor de la información (Ver el apéndice A):

1. *Pre-normalización de scores.* El primer paso consiste en hacer una normalización de scores del tipo *estándar* (Ecuación 2.45, página 36).
2. *Cálculo del valor de la información.* Una vez normalizado el score para todo elemento  $x_i \in Src$  se tiene la expresión  $\overline{Sc}_k(x_i)$  que va entre 0 y 1. Si se divide dicho rango ( $rango \in [0; 1]$ ) en  $p$  partes (donde  $p \in \mathbb{Z}^+$ , se podrá saber el número de documentos  $x_i$  que caen en un segmento  $Seg_k(r)$  (Donde  $Seg_k(r) = \left[ \frac{r-1}{p}; \frac{r}{p} \right]$ ;  $r = 1, 2, \dots, p-1, p$ ;  $\forall k \in Mtd$ ), con lo cual se puede construir un histograma como en la figura 2.9.

Donde el número de elementos que contiene cada segmento  $Seg_r$  está definido como:

$$E(Seg_k(r)) = |\{x_i | \overline{Sc}_k(x_i) \in Seg_r; \forall x_i \in Src\}|; \quad \forall k \in Mtd \quad (2.53)$$

Donde:

$$\sum_{r=1}^p E(Seg_k(r)) = |Src|$$

Luego se podrá calcular la probabilidad que hayan  $E(Seg_k(r)) = d$  documentos en un segmento  $r$  como la frecuencia  $F_k(r)$ :

$$P\left(E(Seg_k(r))=d / \overline{Sc}_k(x_i) \in Seg_k(r)\right) = F_k(r) = \frac{E(Seg_k(r))}{|Src|}; \quad \forall k \in Mtd \quad (2.54)$$

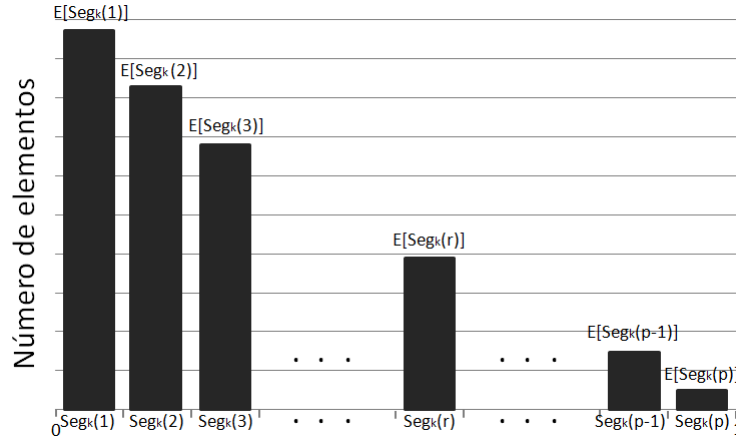


Figura 2.9: Histograma de la linealización de Yu Suzuki et.al. [67]

Con dicho resultado se podrá calcular el *valor de la información* [62]:

$$IV_k(x_i) = -\log_2(F_k(r)); \quad \forall x_i \in Src; \forall k \in Mtd \quad (2.55)$$

Posteriormente, el valor obtenido en la Ecuación 2.55 se deberá linealizar:

$$\overline{IV}_k(x_i) = \frac{IV_k(x_i) - IVMin_k}{IVMax_k - IVMin_k}; \quad x_i \in S_k; \forall k \in Mtd \quad (2.56)$$

Donde:

$$IVMin_k = \{IV_k(x_i) | IV_k(x_i) \leq IV_k(x_j); \forall x_i, x_j \in S_k\} \quad (2.57)$$

$$IVMax_k = \{IV_k(x_i) | IV_k(x_i) \geq IV_k(x_j); \forall x_i, x_j \in S_k\} \quad (2.58)$$

3. *Integración de scores.* El último paso propuesto por Yu Suzuki et. al. [67] es hacer una integración del score linealizado por el método estándar ( $\overline{Sc}_k(x_i)$ ) y el valor de la información ( $IV_k(x_i)$ ), para así finalmente tener el *score linealizado utilizando una técnica de teoría de la información*:

$$\widetilde{Sc}_k(x_i) = \overline{Sc}_k(x_i) \cdot \overline{IV}_k(x_i); \quad \forall x_i \in Src; \forall k \in Mtd \quad (2.59)$$

Hasta este punto, se han presentado una serie de *técnicas de linealización de scores* que se encargan de colocar los scores en un rango que va entre 0 y 1; sin embargo todas las técnicas de linealización  $\overline{Sc}_k(x_i)$  pueden tomar un rango distinto que vaya de 0 a  $C$ , siendo  $C$  una constante ( $C \in \mathbb{R}$ ):

$$\overline{Sc}'_k(x_i) = C [\overline{Sc}_k(x_i)]; \quad C = cte.; C \in \mathbb{R}^+ \quad (2.60)$$

- **Técnicas de combinación** Estas técnicas de combinación de scores fueron propuestas por Fox y Shaw en su trabajo *Combinación de múltiples buscadores*<sup>25</sup>[63].

<sup>25</sup>Combination of multiple searches, por su nombre original en inglés

- **CombMIN.** Esta combinación indica que el score final para un documento  $x_i$  será el score mínimo que entregaron todos los métodos ( $k \in Mtd$ ) para ese elemento.

$$Sc_F(x_i) = \{Sc_k(x_i) | Sc_k(x_i) \leq Sc_l(x_i); \forall k, l \in Mtd\} \quad (2.61)$$

- **CombMAX.** Similar a la técnica anterior, pero en lugar de obtener el score mínimo, se buscará el score máximo que entrega el  $k$ -ésimo método para el elemento  $x_i$ .

$$Sc_F(x_i) = \{Sc_k(x_i) | Sc_k(x_i) \geq Sc_l(x_i); \forall k, l \in Mtd\} \quad (2.62)$$

- **CombMED.** La combinación del valor medio se refiere a encontrar el valor de score promedio que obtuvo un elemento  $x_i$  en todas las  $k$  listas.

$$Sc_F(x_i) = \frac{\sum_{k=1}^{|Mtd|} Sc_k(x_i)}{|Mtd|} \quad (2.63)$$

- **CombSUM.** Esta técnica considera que el valor de score fusionado es igual a la suma de todos los scores obtenidos en los  $k$  métodos para un elemento  $x_i$  en particular.

$$Sc_F(x_i) = \sum_{k=1}^{|Mtd|} \lfloor Sc_k(x_i) \rfloor \quad (2.64)$$

- **CombMNZ.** Proviene de las siglas en inglés de *Mean Non Zero*. La técnica de la CombMNZ utiliza la misma consideración de la *CombSUM* y le agrega el concepto de relevancia ( $R(x_i)$ ) como un producto al valor del *CombSUM*.

$$Sc_F(x_i) = |R(x_i)| \left( \sum_{k=1}^{|Mtd|} \lfloor Sc_k(x_i) \rfloor \right) \quad (2.65)$$

Donde:

$$R(x_i) = |\{x_i | x_i \in S_k; \forall k \in Mtd\}| \quad (2.66)$$

- **CombANZ.** El método *Average Non Zero* utiliza la misma lógica que el *CombMNZ* en el hecho que requiere apoyarse del valor de un conjunto de relevancia, con la diferencia que la relevancia es colocada como divisor del valor *CombSUM*.

$$Sc_F(x_i) = \frac{\sum_{k=1}^{|Mtd|} \lfloor Sc_k(x_i) \rfloor}{|R(x_i)|} \quad (2.67)$$

Finalmente, sea cual sea el método de fusión elegido la lista fusionada  $\rho^F$  tendrá la siguiente forma:

$$\rho^F = \{x_i | Sc_F(x_i) \leq Sc_F(x_j); \{x_i, x_j\} \in |S|\} \quad (2.68)$$

De los métodos de normalización presentados, Yu Suzuki et.al. en [67] demostró que las linealizaciones *estandar*, *SUM* y *ZMUV* no son linealizaciones efectivas. Además Joon Ho Lee en [38] demostró que entre todas las técnicas de combinación propuestas por Aslam y Montague en [49] la única técnica que entrega los mejores resultados es la *CombMNZ*, siendo la que mejor hace recuperación de la información.

## Capítulo 3

# Modelo de Fusión de Datos por Score

En este capítulo se abordará la descripción del modelo de fusión de datos que guía el proyecto de tesis. Como ya se expuso en el capítulo 2, existen dos métodos para fusión de datos: los *métodos de ranking* y los *métodos de score*, donde el uso de estas depende directamente de los valores de salida de cada método de detección de similitud (Ver Figura 2.4, página 22). Si los métodos de detección de similitud entregan resultados como una lista ordenada de elementos (ranking) entonces la única opción será utilizar **métodos de fusión por ranking** (Sección 2.4.2); y si los métodos de detección de similitud entregan resultados como puntaje (score), entonces se aplicarán **métodos de fusión por score** (Sección 2.4.3). También existe un caso, poco común, donde los resultados por *score* se ordenan descendientemente para obtener una lista rankeada de elementos y posteriormente se aplica algún *método de fusión por ranking* sobre esta lista; pero, como lo explican Belkin y Kantor en su trabajo “*Combining the evidence of multiple query representations for information retrieval*” [9] los modelos de combinación por ranking suelen ser menos eficientes que los modelos de Score por ello se recomienda utilizar las combinaciones por score sobre las combinaciones por ranking. Para el caso de este proyecto, los métodos de detección de similitud a usar entregan resultados tipo score entonces se opta por el desarrollo de un *método de fusión de datos por score*.

Se propone el diseño y elaboración de un modelo de fusión de datos de dos etapas: **(a) Linealización de scores** modificando la técnica de Yu Suzuki (Sección 2.4.3. Página 37) la cual se nombró como *Ecuación del valor de la información modificada*; **(b) Método de combinación de scores** basado en la lógica de los métodos presentados por Fox y Shaw [63] (Sección 2.4.3. Página 38) que deonominé *Sistema de combinación geométrico*. Adicionalmente, como parte del proceso general para la fusión de scores se incluye dentro de la *Ecuación del valor de la información modificada* un sistema de *Pesos para los sistemas* detectores de similitud de documentos el cual fue nombrado *Factor de credibilidad* porque serán valores entre cero y uno que indicarán cuánta confianza se le tiene al k-ésimo método de detección de similitud.

### 3.1. Notaciones para el capítulo

Los modelos matemáticos y definiciones que se presentarán a continuación, por equivalencia con el capítulo anterior (Secciones 2.4.2 y 2.4.3), utilizarán las siguientes expresiones y notaciones:

$Spc$	:	Documento sospechoso analizado
$Src$	:	Universo de documentos fuentes para un documento sospechoso. $ Src  = n, \quad n \in \mathbb{Z}^+$
$Mtd$	:	Universo de sistemas de similitud de documentos. $ Mtd  = m, \quad m \in \mathbb{Z}^+$
$k$	:	k-ésimo método de detección de similitud de documentos. $k = 1, 2, 3, \dots, m$
$S_k$	:	Subconjunto de documentos fuentes que entregó el método k-ésimo. $S_k \subseteq Src$
$x_i^k$	:	Elemento i-ésimo del subconjunto $S_k$ . $x_i^k \in S_k$
$Sc_k(x_i)$	:	Score del elemento $x_i$ entregado por el método $k$ .

### 3.2. Consideraciones previas

Todo *Documento Fuente* ( $Spc$ ) y todos los *Documentos Sospechosos* ( $Src$ ) que serán evaluados mediante un sistema de *Detección de Similitud* podrán subdividirse en partes menores. Es decir, la comparación de similitud no sólo podrá ser por la comparación de *Documento Completo vs. Documento Completo*, sino también entre sus partes como *Párrafo i del Documento Sospechoso vs. Párrafo j del Documento Fuente* o *Frase  $f_p$  del párrafo i del Documento Sospechoso vs. Frase  $f_q$  del párrafo j del Documento Fuente*, entre otras<sup>1</sup>. Para el caso de los métodos de *Detección de Similitud de Documentos* que entreguen resultados por que entregue resultados por Score ( $Sc_k(x_i)$ ) con la división mencionada se podrá obtener una matriz ( $M$ ) de datos cuyas dimensiones dependerán directamente de que tan “fina” se desee hacer la evaluación entre Documentos<sup>2</sup>.

La figura 3.1 muestra un ejemplo de una *Matriz de Scores* con valores luego de la comparación tipo *Página i del Documento Sospechoso vs. Página j del Documento Fuente* quedando una matriz de tres dimensiones:

$$M = [Spc(Pag_p), Src_i(Pag_q), Mtd_k] \quad \forall i \in Src, k \in Mtd$$

El *Documento Sospechoso* (Figura 3.1), se encuentra en la *Dimensión 1*. Mientras que los *Documentos Fuentes* están ubicados en la *Dimensión 2* y los *Métodos de Detección de similitud* están en la *Dimensión 3*. Además los *Documentos Fuente* y *Sospechosos* no necesariamente deben contar con la misma estructura, en este caso con el mismo número de páginas, el *Documento Sospechoso* posee tres páginas al igual que los *Documentos Fuentes 1 y 2* pero el *Documento Fuente 3* sólo tiene dos. Esto es posible porque la lógica de

<sup>1</sup>Las particiones más comunes para un documento literal son: Documento completo, Páginas, Párrafos, Frases, N-grama, Palabra.

<sup>2</sup>El nivel de detalle también dependerá del *Método de Detección de Similitud de Documentos*. Existen métodos que admiten un análisis hasta por palabra, mientras que otros lo hacen directamente por Documento



		Documento Fuente 1 (Src 1)			Documento Fuente 2 (Src 2)			Documento Fuente 3 (Src 3)	
		Pág. 1	Pág. 2	Pág. 3	Pág. 1	Pág. 2	Pág. 3	Pág. 1	Pág. 2
Documento Sospechoso (SpC)	Pág. 1	0.74	0.11	0.09	0.21	0.17	0.05	0.82	0.31
	Pág. 2	0.13	0.28	0.15	0.04	0.93	0.02	0.31	0.10
	Pág. 3	0.04	0.02	0.01	0.31	0.06	0.98	0.02	0.08

Figura 3.1: Matriz de Scores - Detalle por página

funcionamiento de cualquier *Método de Detección de Similitud* es independiente a las características externas del documento a comparar; el método sólo se encarga de reconocer la similitud de los *Textos A* y *B* y le resulta “indiferente” que dichos *Textos* sean párrafos, frases o documentos completos. La Figura 3.2 explica mejor este funcionamiento, el *Método de Detección de Similitud* toma como entrada (Input) dos textos los procesa y entrega un resultado de la similitud entre ellos.

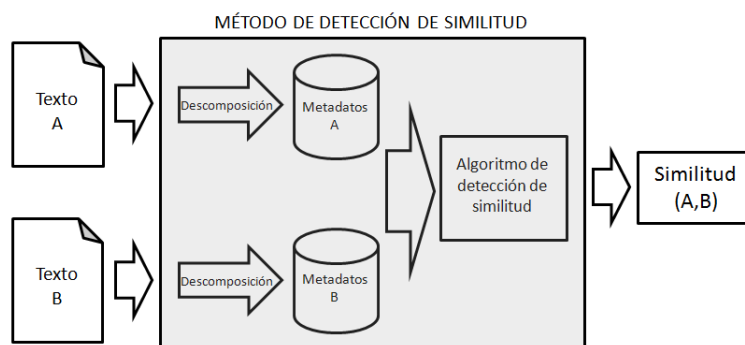


Figura 3.2: Método de Detección de Similitud

Entonces, como ya se expuso, el nivel de detalle en la comparación de Documentos dependerá del modo de ingreso de los textos por ello el *Modelo de Fusión de Datos* que se describe en la Sección 3.6 (Página 50) también acepta esta “finura” en las comparaciones. Sin embargo, para que las expresiones posteriores sean más cómodas de entender y analizar en esta tesis, en adelante sólo se considerarán comparaciones entre *Documentos completos* (Documento A vs. Documento B). Esta consideración hacen que nuestra Matriz *M* de la Figura 3.1 se reduzca a la Matriz *M* de la Figura 3.3.

En la Figura 3.3 (a), se tiene a la Matriz *M* que aún conserva sus tres dimensiones a pesar de la reducción de datos, pero cómo sólo existe un único *Documento Sospechoso* el número de elementos en la Dimensión 1 siempre será único e igual a *SpC*. Si se elimina la Dimensión 1 y se dejan la Dimension 2

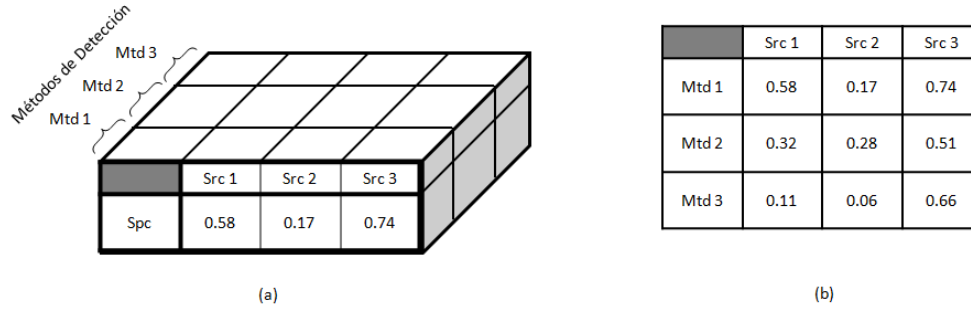


Figura 3.3: Matriz de Scores - Por documento completo

(Documentos Fuentes) y Dimensión 3 (Métodos de Detección de Similitud), la Matriz  $M$  cambiará a dos Dimensiones como en la Figura 3.3 (b). Donde:

$$M = [Mtd_k, Src_i] \quad \forall k \in Mtd, i \in Src$$

La Matriz  $M$  en dos dimensiones, con las expresiones de la Sección 3.1 (Página 41), se podrá generalizar para un único *Documento Sospechoso* como en el Cuadro 3.1.

	$Src_1$	$Src_2$	$Src_3$	$\dots$	$Src_i$	$\dots$	$Src_{n-1}$	$Src_n$
$Mtd_1$	$Sc_1(x_1)$	$Sc_1(x_2)$	$Sc_1(x_3)$	$\dots$	$Sc_1(x_i)$	$\dots$	$Sc_1(x_{n-1})$	$Sc_1(x_n)$
$Mtd_2$	$Sc_2(x_1)$	$Sc_2(x_2)$	$Sc_2(x_3)$	$\dots$	$Sc_2(x_i)$	$\dots$	$Sc_2(x_{n-1})$	$Sc_2(x_n)$
$Mtd_3$	$Sc_3(x_1)$	$Sc_3(x_2)$	$Sc_3(x_3)$	$\dots$	$Sc_3(x_i)$	$\dots$	$Sc_3(x_{n-1})$	$Sc_3(x_n)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
$Mtd_k$	$Sc_k(x_1)$	$Sc_k(x_2)$	$Sc_k(x_3)$	$\dots$	$Sc_k(x_i)$	$\dots$	$Sc_k(x_{n-1})$	$Sc_k(x_n)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
$Mtd_{m-1}$	$Sc_{m-1}(x_1)$	$Sc_{m-1}(x_2)$	$Sc_{m-1}(x_3)$	$\dots$	$Sc_{m-1}(x_i)$	$\dots$	$Sc_{m-1}(x_{n-1})$	$Sc_{m-1}(x_n)$
$Mtd_m$	$Sc_m(x_1)$	$Sc_m(x_2)$	$Sc_m(x_3)$	$\dots$	$Sc_m(x_i)$	$\dots$	$Sc_m(x_{n-1})$	$Sc_m(x_n)$

Cuadro 3.1: Matriz genérica de Scores

Más adelante, en la Sección 3.6 se presenta el *Modelo de Fusión de Datos por Score* que busca integrar todos los resultados de los Métodos ( $Mtd$ ) de Detección de Similitud de la Matriz  $M$  (Ver Cuadro 3.1) para finalmente tener un único vector de Scores respecto al *Documento Sospechoso* y los *Documentos Fuentes* como en el Cuadro 3.2

	$Src_1$	$Src_2$	$Src_3$	$\dots$	$Src_i$	$\dots$	$Src_{n-1}$	$Src_n$
$Mtd_F$	$Sc_F(x_1)$	$Sc_F(x_2)$	$Sc_F(x_3)$	$\dots$	$Sc_F(x_i)$	$\dots$	$Sc_F(x_{n-1})$	$Sc_F(x_n)$

Cuadro 3.2: Vector de Scores Fusionados

Donde:

$Sc_F(x_i)$  : Score del elemento  $x_i$  resultante luego del análisis por el *Método de Fusión de Datos por Scores*.

Finalmente con los resultados del Vector del Cuadro 3.2 se podrá ordenar descendentemente por Score, además reconocer cuál *Documento Fuente* tiene más similitud con el *Documento Sospechoso* y por lo tanto mayor probabilidad que haya copia entre ellos.

En las siguientes secciones se expondrán las partes relevantes que fueron motivo de la investigación para el proyecto de tesis: **(a) La Ecuación del Valor de la Información Modificada.** Que incluye el análisis y modificación a la ecuación en el método de linealización de Scores propuesta por Yu Suzuki [67]. **(b) Sistema de Combinación Geométrica.** Una variación de los métodos de combinación que propusieron Fox y Shaw [63] que busca “castigar” el *Score Fusión* de los métodos que entregaron bajos puntajes de similitud entre documentos. **(c) Factor de Credibilidad.** Una variable que indicará la *Confianza* que se tiene ante la detección de similitud de un método (*Mtd*) en particular, esta variable será ingresada por el usuario y se usará para “calibrar” el *Modelo de Fusión de Datos por Score* para que el “castigo” dado por la *Combinación Geométrica* no sea tan drástico.

### 3.3. Ecuación del Valor de la Información Modificada

De acuerdo a las técnicas de linealización de score presentados en la sección 2.4.3 (Página 36) y conforme al análisis numérico del Apéndice C, el método más eficiente para linealización es el propuesto por Yu Suzuki et. al., dicho método incluye conceptos de entropía en teoría de la información (Ver Apéndice A) para definir qué tan valioso es el score de un documento respecto a los demás sospechosos.

El objetivo de esta sección no es ahondar en detalles de la ecuación de Suzuki et. al. [67] que ya se encuentran explicados en la Sección 2.4.3, sin embargo es importante resaltar de las tres etapas propuestas por Suzuki: (a) *Linealización de scores*, (b) *Cálculo del Valor de la Información* y (c) *Integración de scores*, la propuesta de esta tesis es modificar la segunda etapa (*Cálculo del valor de la información*) con el fin de generalizar la técnica de Suzuki. La modificación propuesta quedaría expresada como:

$$IV_k(x_i) = \underbrace{-\log_2(F_k(r))}_{YuSuzuki\ et.\ al} + \underbrace{\left(\frac{1}{F_k(r)}\right) \log_2 |Mtd|}_{Propuesta}; \quad \forall x_i \in Src; \forall k \in Mtd \quad (3.1)$$

Con esta propuesta se conseguirá que luego de evaluar documentos de *mejor score* en la **Ecuación del Valor de la Información Modificada** estos obtengan un mayor puntaje que aquellos de *peor score*. Además, como depende del número de métodos de similitud de documentos ( $Mtd$ ), el puntaje relativo será mayor para los documentos de *mejor score*. Estas características permitirán que la *Ecuación del valor de la información modificada* entregue un resultado valioso, que resalta a los mejores documentos, haciendo que el siguiente paso de la fusión de scores (Combinación de Scores. Sección 3.4) sea más eficiente.

En el párrafo anterior se hizo clara mención de la definición *mejor* y *peor score*, porque la calificación de *mejor* o *peor* es relativa al contexto en el cual se trabaje y depende directamente del tipo de resultados que otorguen los métodos a fusionar. En este caso particular un *mejor score* será igual a un *alto score*, y contrariamente un *peor score* será igual a un *bajo score*. De un universo de  $Src$  documentos fuentes para un único documento sospechoso, y considerando que  $Src$  es un valor representativo, muy pocos documentos fuentes evaluados por algún  $k$ -ésimo método de detección de similitud ( $x_i^k$ ) tendrán un *alto score* ( $Sc_k(x_i) \rightarrow 1$ ), es decir tendrán una alta similitud con el documento sospechoso, por lo tanto serán *mejores* respecto a los otros documentos fuentes ( $Sc_k(x_j) \rightarrow 0; \forall j \neq i$ ) para algún posterior análisis.

Como se afirmó anteriormente, la propuesta de la **Ecuación del Valor de la Información Modificada** es una generalización del método de Yu Suzuki et. al. porque si la Ecuación 3.1 se analiza para un único método de detección de copia, o sea  $|Mtd| = 1$ , la modificación propuesta toma el valor cero, que en efectos de suma es un valor nulo:

$$Si : \quad |Mtd| = 1; \quad \Rightarrow \quad \left(\frac{1}{F_k(r)}\right) \log_2 |Mtd| = 0$$

Por lo tanto la expresión del Valor de la Información propuesta por Suzuki et. al. [67] queda intacta e igual a la Ecuación 2.55 presentada en la sección anterior (Página 38).

$$IV_k(x_i) = -\log_2(F_k(r)); \quad \forall x_i \in Src; \forall k \in Mtd \quad (3.2)$$

Queda bien aclarar que la **Ecuación del Valor de la Información Modificada** (Ecuación 3.1) se torna relevante cuando existe fusión de métodos. Es decir, que el número de métodos de detección de similitud a fusionar sea mayor a 1 ( $|Mtd| > 1$ ) como es lógico.

### 3.4. Sistema de Combinación Geométrica

Para la combinación de datos se utilizará una variación a los métodos presentados por Fox y Shaw [63] y que fueron repasados en el capítulo anterior (Sección 2.4.3, Página 38) al que se denominó **Sistema de combinación geométrica** porque para combinar utiliza el producto de scores de los documentos analizados con el k-ésimo método de similitud y luego saca la raíz k-ésima, lo que es similar a aplicar la media geométrica para los scores de un documento analizado por k métodos de similitud ( $k \in Mtd$ ).

$$Sc_F(x_i) = \sqrt[|Mtd|]{\prod_{k=1}^{|Mtd|} Sc_k(x_i)} \quad (3.3)$$

Se hace esta propuesta porque luego de la linealización lograda con la **Ecuación del Valor de la Información Modificada**, donde se resaltan con un alto puntaje aquellos documentos con *mejor score*, el **Sistema de Combinación Geométrica** luego de la fusión de scores asignará un alto puntaje a aquellos documentos que cuenten con un alto o mediano score en la mayoría de métodos de similitud. Es decir, el *Sistema de Combinación Geométrica* castigará aquellos documentos que tengan bajo score y a los documentos que hayan sido detectados con alta similitud por unos pocos detectores k ( $k \in Mtd$ ) pero en general la mayoría de detectores de similitud no lo califican con alta probabilidad de copia.

Por ejemplo, asumamos que se cuenta con 6 distintos *métodos de detección de similitud* los cuales se utilizaron para analizar un único *documento sospechoso* contra cinco *documentos fuentes A, B, C, D y E*. Luego de la linealización con la *Ecuación del Valor de la Información Modificada* se obtuvieron los siguientes resultados:

$Sc_1(A)=0,97$	$Sc_1(B)=0,02$	$Sc_1(C)=0,10$	$Sc_1(D)=0,51$	$Sc_1(E)=0,96$
$Sc_2(A)=0,89$	$Sc_2(B)=0,07$	$Sc_2(C)=0,86$	$Sc_2(D)=0,46$	$Sc_2(E)=0,99$
$Sc_3(A)=0,99$	$Sc_3(B)=0,11$	$Sc_3(C)=0,04$	$Sc_3(D)=0,63$	$Sc_3(E)=0,03$
$Sc_4(A)=0,92$	$Sc_4(B)=0,03$	$Sc_4(C)=0,06$	$Sc_4(D)=0,50$	$Sc_4(E)=0,91$
$Sc_5(A)=0,95$	$Sc_5(B)=0,04$	$Sc_5(C)=0,91$	$Sc_5(D)=0,59$	$Sc_5(E)=0,98$
$Sc_6(A)=0,88$	$Sc_6(B)=0,08$	$Sc_6(C)=0,14$	$Sc_6(D)=0,48$	$Sc_6(E)=0,99$

Se observa que para cada tipo de *Documento Fuente*, los métodos de detección de similitud entregaron distintos resultados: El *Documento Fuente A* tiene una alta similitud con el *Documento Sospechoso* pues todos los métodos de similitud dieron un valor cercano a uno<sup>3</sup>. Para el *Documento Fuente B* ocurre el caso contrario que en A, los métodos indicaron una baja similitud. El *Documento Fuente C*, obtuvo resultados bastante peculiares, para los detectores de similitud 2 y 5 hay mucho parecido entre el par Sospechoso–Fuente y para los detectores de similitud 1, 3, 4 6 no hay parecido. El *Documento Fuente D* tiene valores bastante regulares, todos los detectores de similitud lo posicionaron con una cercanía media (alrededor de 0,50). Finalmente el *Documento Fuente E* entregó un caso parecido al *Documento Fuente C* con la diferen-

<sup>3</sup>Después de la linealización, los scores se encuentran en el rango entre 0 y 1 ( $rango \in [0; 1]$ ). Siendo 0 poca similitud entre documentos y 1 alta similitud.

cia que sólo el detector 3 indicó poca similitud, mientras que los demás indicaron alta similitud entre el par Sospechoso–Fuente.

Ahora, si utilizamos la *Ecuación de Combinación Geométrica* propuesta (Ecuación 3.16), se obtendrán resultados bastante interesantes<sup>4</sup>:

$$S_{c_F}(A)=0,93 \quad S_{c_F}(B)=0,05 \quad S_{c_F}(C)=0,17 \quad S_{c_F}(D)=0,52 \quad S_{c_F}(E)=0,54$$

Para el *Documento Fuente A* la alta similitud con el par *Documento Sospechoso–Fuente* se mantiene, lo contrario con el *Documento Fuente B*. El score del *Documento Fuente C* se observa que baja mucho porque muchos de los métodos indicaron que no existía similitud. El caso del *Documento Fuente D* los valores siguen estables, igual como ocurrió con los Documentos A y B. El *Documento Fuente E* tiene un resultado de fusión importante, a pesar que muchos detectores indicaron similitud entre Sospechoso–Fuente bastó que uno entregue un bajo valor para que el score fusionado baje hasta parecerse al caso del *Documento Fuente D* que tienen un puntaje medio.

Es debido a esta característica de “castigar” el *Score Fusionado* que se elige utilizar el **Sistema de Combinación Geométrica**, porque será un modo más de resaltar los Documentos Fuentes más parecidos al Documento Sospechoso sobre cualquier otro Documento Fuente.

Cabe destacar que para utilizar este método, los scores después de la linealización nunca deben ser cero porque anularía el resultado de la combinación. Por lo tanto es importante que el rango de scores luego de la linealización con la *Ecuación del Valor de la Información Modificada* se encuentren entre  $c$  y 1 ( $rango \in [c; 1]$ ) donde  $c$  es un valor que tiende a cero, pero sin llegar a ser cero ( $c \rightarrow 0$ ).

---

<sup>4</sup>Nota: Todos los resultados fueron redondeados a dos dígitos decimales

### 3.5. Factor de Credibilidad

La definición del *Factor de Credibilidad* está bastante ligada al *Sistema de Combinación Geométrica* descrito anteriormente. Como ningún método de detección de similitud entre documentos posee eficacia total en todos los casos de análisis, se hace necesario un factor de corrección que en este contexto fue definido como *Factor de Credibilidad* porque indicará que tanta credibilidad le otorga el usuario a cada método de similitud entre documentos que está fusionando. Este Factor es una variable modificable por el usuario que mediante juicio experto, para cada contexto distinto, podrá indicar qué método de detección de similitud es mejor que otro. Por ejemplo, en el contexto de la detección de plagio, existirán métodos de similitud de documentos Fuentes–Sospechoso que funcionen mejor ante textos muy extensos, mientras que otros lo serán para textos breves; si al momento de analizar una dupla *Fuente-Sospechoso* que contiene texto extenso será recomendable agregarle un mayor *Factor de Credibilidad* a los métodos que funcionan mejor con grandes textos.

El *Factor de Credibilidad* afectará directamente el valor del Score entregado por algún k-ésimo método de fusión durante la aplicación del *Sistema de Combinación Geométrica*, quedando la siguiente expresión:

$$Sc_F(x_i) = \sqrt[|Mtd|]{\prod_{k=1}^{|Mtd|} [Sc_k(x_i)]^{FC_k}} \quad (3.4)$$

Donde:

$FC_k$  : Factor de Credibilidad de k-ésimo método de detección de similitud entre documentos.  
 $FC_k \in [0; 1]$

El valor del *Factor de Credibilidad* tomará valores entre 0 y 1, donde un valor de  $FC_k \rightarrow 0$  indicará que el sistema tiene poca credibilidad y un valor de  $FC_k \rightarrow 1$  indicará que el sistema es bastante confiable para esa fusión.



### 3.6. Modelo de Fusión de Datos por Score

Esta sección describe la propuesta del **Modelo de Fusión de Datos por Score** el cual integra todos los aportes de este proyecto de tesis en un algoritmo de cinco pasos <sup>5</sup>

Para comprender mejor el proceso de Fusión de Datos por Score, apoyémonos en el gráfico presentado en el Capítulo 2:

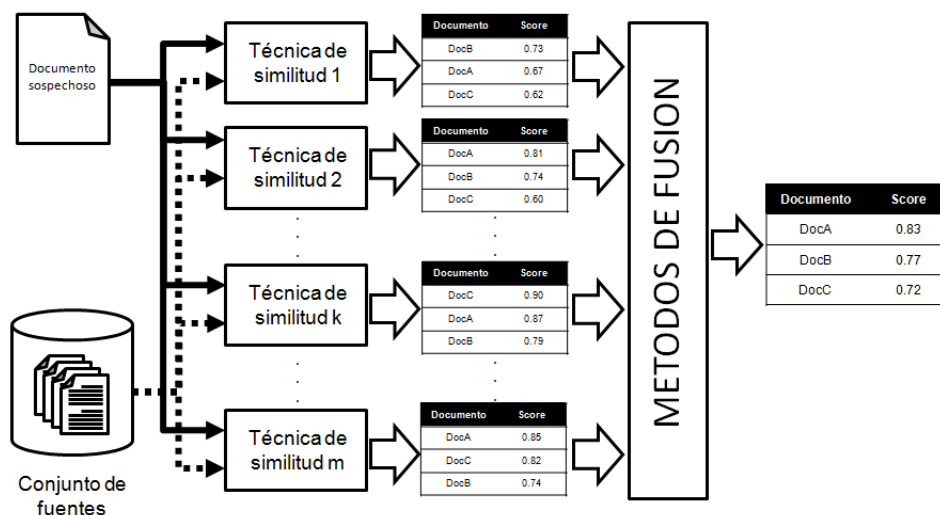


Figura 3.4: Sistemas de fusión de datos

Donde se tiene un *Documento Sospechoso* (*Spc*) que será comparado contra un conjunto de *Documentos Fuentes* (*Src*) por cada sistema detector de plagio, todos estos documentos entran en duplas *Sospechoso–Fuente* para ser comparados por las diversas *Técnicas de similitud*, posterior al proceso de comparación de similitud cada técnica entregará un *Score* para el par *Sospechoso–Fuente* analizado. Si este análisis se repite con todos los *Documentos Fuentes* del conjunto, se tendrá un vector de resultados con un listado de todos los *Documentos Fuentes* junto a su respectivo *Score*; finalmente serán todos los vector entregados por las distintas técnica de similitud los que entrarán al *Sistema de Fusión de Datos* el cual finalmente producirá el vector de Scores fusionados.

Las siguiente cinco fases se enmarcan dentro de la Etapa de *Fusión de Datos* que en la figura está nombrada como *Métodos de Fusión*:

- **FASE 1: Linealización Estándar de Scores**

Los Scores en los vectores de resultados que salen de las *Técnicas de Similitud de Documentos* no necesariamente tienen un rango definido. En esta etapa se busca que los Scores del vector tengan

<sup>5</sup>Se prefirió utilizar cinco fases porque se considera una mejor división que sólo las dos etapas de los modelos de fusión de Datos: (a) Linealización de Scores y (b) Combinación de Scores

un rango definido que va de cero a uno, siendo cero la peor similitud del *Documento Fuente con el Sospechoso* y uno la mejor similitud.

$$\overline{Sc}_k(x_i) = \frac{Sc_k(x_i) - ScMin_k}{ScMax_k - ScMin_k}; \quad x_i \in S_k; \forall k \in Mtd \quad (3.5)$$

$$\overline{Sc}_k(x_i) \in [0; 1]$$

Donde:

$$ScMin_k = \{Sc_k(x_i) | Sc_k(x_i) \leq Sc_k(x_j); \forall x_i, x_j \in S_k\} \quad (3.6)$$

$$ScMax_k = \{Sc_k(x_i) | Sc_k(x_i) \geq Sc_k(x_j); \forall x_i, x_j \in S_k\} \quad (3.7)$$

■ **FASE 2: Cálculo de la Frecuencia de Scores**

La Fase del Cálculo de la Frecuencia de Scores, es que a partir de los datos se pueda conocer con qué frecuencia los puntajes del Vector de Scores caen en un segmento determinado dentro del rango de cero a uno. Por ejemplo la figura:

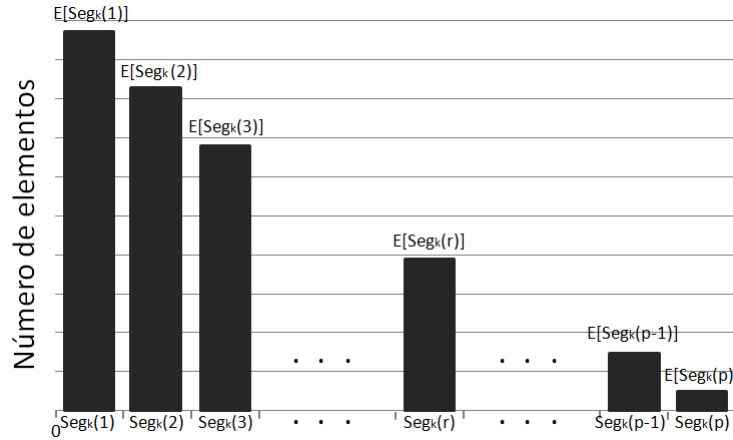


Figura 3.5: Frecuencia que los Scores se encuentren en un Segmento r-ésimo

Se observa que hay pocos Scores altos, es decir que hay pocos *Documentos Fuentes* que tienen alta similitud con el *Documento Sospechoso*. Esto es bastante lógico porque al momento de comparar el *Documento Sospechoso* contra un gran conjunto de *Documentos Fuentes* muy pocos tendrán gran similitud, mientras que la mayoría tendrá muy poca o mediana similitud.

Para construir la Figura 3.5 primero el rango de cero a uno, se dividió en  $p$ -segmentos, y luego se construyó el histograma que contaba el número de *Scores de Documentos*  $E(Seg_k(r))$  que pertenecía a algún segmento r-ésimo  $Seg_k(r)$ :

$$E(Seg_k(r)) = |\{x_i | \overline{Sc}_k(x_i) \in Seg_r; \forall x_i \in Src\}|; \quad \forall k \in Mtd \quad (3.8)$$

Donde:

$$Seg_k(r) = \left[ \frac{r-1}{p}; \frac{r}{p} \right]; r = 1, 2, \dots, p-1, p; \forall k \in Mtd \quad (3.9)$$

$p$  : Número de divisiones que se requiere repartir el *rancho*  $\in [0; 1]$ ;  $p \in \mathbb{Z}^+$

Luego de conocer el valor  $E(Seg_k(r)) = d$  se puede calcular la Frecuencia  $F_k(r)$  para el segmento  $r$ -ésimo mediante la siguiente expresión:

$$P\left(\frac{E(Seg_k(r))=d}{\overline{Sc_k(x_i) \in Seg_k(r)}}\right) = F_k(r) = \frac{E(Seg_k(r))}{|Src|} = \frac{d}{|Src|}; \quad \forall k \in Mtd \quad (3.10)$$

■ **FASE 3: Cálculo del Valor de la Información Modificada**

Con la Frecuencia  $F_k(r)$  calculada y conociendo el número de *Métodos de Detección de Similitud de Documentos* ( $Mtd$ ) a fusionar, se podrá aplicar el *Cálculo del Valor de la Información Modificada* (Sección 3.3).

$$IV_k(x_i) = -\log_2(F_k(r)) + \left(\frac{1}{F_k(r)}\right) \log_2|Mtd|; \quad \forall x_i \in Src; \forall k \in Mtd \quad (3.11)$$

Los valores obtenidos en  $IV_k(x_i)$  pueden ser del rango cero a  $\mathbb{R}^+$ , por ello se les linealiza para tener valores entre cero y uno.

$$\overline{IV}_k(x_i) = \frac{IV_k(x_i) - IVMin_k}{IVMax_k - IVMin_k}; \quad x_i \in S_k; \forall k \in Mtd \quad (3.12)$$

Donde:

$$IVMin_k = \{IV_k(x_i) | IV_k(x_i) \leq IV_k(x_j); \forall x_i, x_j \in S_k\} \quad (3.13)$$

$$IVMax_k = \{IV_k(x_i) | IV_k(x_i) \geq IV_k(x_j); \forall x_i, x_j \in S_k\} \quad (3.14)$$

■ **FASE 4: Integración del Score Linealizado y del Valor de la Información Modificada**

Esta Fase integra el *Score Linealizado* obtenido en la Fase 1 con el resultado del *Valor de la Información Modificada* de la Fase 3 mediante un producto punto.

$$\widetilde{Sc}_k(x_i) = \overline{Sc}_k(x_i) \cdot \overline{IV}_k(x_i); \quad \forall x_i \in Src; \forall k \in Mtd \quad (3.15)$$

■ **FASE 5: Combinación geométrica**

El paso final del *Sistema de Fusión de Scores* es combinar todos los *Scores* obtenidos en la Fase 4, esto se realiza con el *Sistema de Combinación Geométrica* (Sección 3.4) que incluye el *Factor de Credibilidad* (Sección 3.5) de las distintas *Técnicas de Detección de Similitud*

$$Sc_F(x_i) = \sqrt[|Mtd|]{\prod_{k=1}^{|Mtd|} [\widetilde{Sc}_k(x_i)]^{FC_k}} \quad (3.16)$$

Donde:

$FC_k$  : Factor de Credibilidad de k-ésimo método de detección de similitud entre documentos.  
 $FC_k \in [0; 1]$

### Modelo de Fusión de Datos integrado

A nivel matemático, si se integran las cinco fases del *Modelo de Fusión de Datos por Score* en una sola expresión para no tener que realizar los cinco pasos, la expresión queda como en la Ecuación 3.17

$$Sc_F(x_i) = \sqrt[|Mtd|]{\prod_{k=1}^{|Mtd|} \left[ \frac{Sc_k(x_i) - ScMin_k}{ScMax_k - ScMin_k} \cdot \frac{-\log_2(F_k(r)) + \left(\frac{1}{F_k(r)}\right) \log_2 |Mtd| - IVMin_k}{IVMax_k - IVMin_k} \right]^{FC_k}} \quad (3.17)$$

La cual, debido al logaritmo que incluye, puede modificarse para hacer que la expresión sea un poco más compacta. Quedando:

$$Sc_F(x_i) = \sqrt[|Mtd|]{\prod_{k=1}^{|Mtd|} \left[ \log_2 \left( \left[ \frac{IVMax_k - IVMin_k \sqrt{\frac{F_k(r) \sqrt{|Mtd|}}{F_k(r) \cdot 2^{(IVMin_k)}}}}{\left(\frac{Sc_k(x_i) - ScMin_k}{ScMax_k - ScMin_k}\right)} \right]^{FC_k} \right)} \quad (3.18)$$

Además, para efectos prácticos, se puede considerar el resultado en escala logarítmica consiguiendo evitar de ese modo a la pitatoria que es reemplazada por una sumatoria. La siguiente expresión muestra el mencionado cambio, se utilizó una escala logarítmica base 10 por ser un valor que ayuda a manejar mejor los resultados:

$$\log_{10} [Sc_F(x_i)] = \frac{1}{|Mtd|} \cdot \sum_{k=1}^{|Mtd|} \left\{ FC_k \cdot \log_{10} \left[ \log_2 \left( \left[ \frac{IVMax_k - IVMin_k \sqrt{\frac{F_k(r) \sqrt{|Mtd|}}{F_k(r) \cdot 2^{(IVMin_k)}}}}{\left(\frac{Sc_k(x_i) - ScMin_k}{ScMax_k - ScMin_k}\right)} \right]^{FC_k} \right) \right] \right\} \quad (3.19)$$

### 3.7. Algoritmo del Modelo de Fusión de Datos por Score

Por la construcción del Modelo de Fusión de Datos (Sección 3.6. Página 50) todas las fases fueron diseñadas para que posteriormente se puedan implementar en algún lenguaje de programación<sup>6</sup>. En esta sección se explican los algoritmos que integran las ecuaciones del *Modelo de Fusión de Datos* y las lógicas de implementación; a diferencia de la Sección 3.6 el *Algoritmo del Modelo de Fusión de Datos* consta de 6 Fases que van desde la Fase 0 a la Fase 5:

- Fase 0** : Lectura de Scores entregados por los métodos *Mtd* y la carga de la Matriz *M*.
- Fase 1** : Linealización inicial de los Scores de la Matriz *M*.
- Fase 2** : Cálculo de la Frecuencia de Scores.
- Fase 3** : Cálculo del Valor de la Información.
- Fase 4** : Integración de los Scores Linealizados (Fase 1) y del Valor de la Información (Fase 3).
- Fase 5** : Fusión de Scores mediante la Combinación Geométrica, incluye el Factor de credibilidad.

A continuación se abordarán con más detalle cada una de las mencionadas fases del *Algoritmo de Fusión de Datos por Score*. Se recomienda seguir la lógica de los siguientes algoritmos para la implementación del *Modelo de Fusión de Datos* presentando en la Sección 3.6, también es bueno recalcar que el objetivo de esta tesis no es desarrollar un software de Fusión de datos por dicho motivo los algoritmos presentados no se encuentran optimizados por lo que admiten alguna posterior revisión y mejora.

#### ■ FASE 0: Lectura de Scores y carga de matriz inicial

Como se explicó en la Sección de Definiciones previas (Sección 3.2) con los Scores obtenidos previamente por el análisis de similitud entre Documento Sospechoso vs. Documentos Fuentes por los diversos Métodos de Detección de Similitud ( $Sc_k(x_i)$ ) se puede generar una Matriz *M*, a la que a partir de ahora llamaremos *Matriz de Scores*, de dos Dimensiones que contenga todos los resultados (Ver Cuadro 3.1). El algoritmo será:

---

#### Algorithm 3.7.1: Fase 0

---

**Data:**  $Mtd, Src, Sc_k(x_i) \forall k \in Mtd, x_i \in Src$

**Result:** *M*: Matriz de Scores

- 1 Inicializar  $M = Dim [|Mtd|, |Src|]$ ;
  - 2 **foreach**  $k \in Mtd$  **do**
  - 3     **foreach**  $x_i \in Src$  **do**
  - 4          $M(k, i) = Sc_k(x_i)$ ;
  - 5 **return** *M* ;
- 

<sup>6</sup>Para la tesis se eligió el software libre *Java*. Ver Capítulo 4

- **FASE 1: Linealización inicial de Scores** Los Scores cargados en la *Matriz de Scores*  $M$  no están normalizados y pueden contener rangos en cualquier escala real ( $Rango \in [a, b]$ , donde  $a, b \in \mathbb{R}$ ), se les linealiza para tenerlos en un rango estándar de 0 a 1 ( $Rango \in [0, 1]$ )<sup>7</sup>

---

**Algorithm 3.7.2:** Fase 1

---

**Data:**  $M$ : Matriz de Scores  
**Result:**  $M$ : Matriz de Scores normalizada

```

1 Inicializa  $|Mtd|$ :  $|Mtd| = NumCol(M)$ ;
2 Inicializa  $|Src|$ :  $|Src| = NumFil(M)$ ;
3 foreach  $k \in Mtd$  do
4      $max = -\infty$ ;
5      $min = +\infty$ ;
6     foreach  $x_i \in Src$  do
7         if  $M(k, i) > max$  then
8              $max = M(k, i)$ ;
9         if  $M(k, i) < min$  then
10             $min = M(k, i)$ ;
11    if  $max \neq min$  then
12        foreach  $x_i \in Src$  do
13            if  $M(k, i) \neq min$  then
14                 $M(k, i) = \frac{M(k, i) - min}{max - min}$ ;
15            else
16                 $M(k, i) \rightarrow 0$ ;
17    else
18        foreach  $x_i \in Src$  do
19             $M(k, i) = -1$ 
20 return  $M$ ;

```

---

El Algoritmo 3.7.2 consta de dos etapas: La primera etapa que por cada *Método de Detección de Similitud* ( $Mtd$ ) encuentra el *Máximo Score* ( $max$ ) y el *Mínimo Score* ( $min$ ) entre los resultados Documento Sospechoso vs. Documentos Fuentes entregados.

La segunda etapa es la asignación del *Score linealizado* que se subdivide en tres partes: **(a) Caso típico distinto del mínimo.** El score se linealiza mediante la expresión presentada en la Ecuación 3.5. **(b) Caso típico igual al mínimo.** Cuando el Score a linealizar es igual al mínimo, si se aplica la Ecuación 3.5 el valor será igual a cero. Sin embargo como se explicó al final de la Sección 3.4 este valor no debe ser cero sino un valor que *tienda a cero*. **(c) Caso atípico.** Este caso sólo se cumple cuando todos los Scores del método  $Mtd$  tienen el mismo valor. En estos casos si se aplica la Ecuación 3.5 el valor de la linealización resulta indefinido; para evitar este resultado se opta por colocar el valor  $-1$  que es fácilmente reconocible y escapa del rango entre cero y uno con el que se desea linealizar.

---

<sup>7</sup>Se considera que todos los Métodos de Detección de Similitud entregan un bajo score cuando no hay similitud entre Documentos, y viceversa cuando hay mucha similitud.

- **FASE 2: Cálculo de la Frecuencia de Scores** Para hallar la Frecuencia de los Scores, se requiere particionar el rango de 0 a 1 en  $p$ -segmentos, donde  $p$  es un valor ingresado a criterio del usuario; el valor de  $p$  influirá en los resultados finales por ello más adelante en el Capítulo 5 se realiza un profundo análisis de la correcta calibración de este valor.

Esta Fase busca crear para cada  $Mtd$  una tabla con las siguientes características<sup>8</sup>:

Segmento	1	2	...	$r$	...	$p - 1$	$p$
Cantidad	$E(Seg_k(1))$	$E(Seg_k(2))$	...	$E(Seg_k(r))$	...	$E(Seg_k(p - 1))$	$E(Seg_k(p))$
Frecuencia	$F_k(1)$	$F_k(2)$	...	$F_k(r)$	...	$F_k(p - 1)$	$F_k(p)$

Cuadro 3.3: Vector de Frecuencias de Scores

Se tendrá un algoritmo para el cálculo de la *Cantidad* de elementos que pertenecen a un segmento  $r$ -ésimo en particular y otro algoritmo para el cálculo de la *Frecuencia*.

---

**Algorithm 3.7.3:** Fase 2: Cantidad de elementos

---

**Data:**  $p, M$ : Matriz de Scores normalizada

**Result:**  $Elementos_r, \forall r \in p$

```

1 Inicializa  $|Mtd|$ :  $|Mtd| = NumCol(M)$ ;
2 Inicializa  $|Src|$ :  $|Src| = NumFil(M)$ ;
3 Inicializa  $Elementos_r = 0, \forall r \in p$ ;
4 foreach  $k \in Mtd$  do
5   | foreach  $x_i \in Src$  do
6     | | foreach  $r \in p$  do
7       | | | if  $M(k, i) \in \left[ \frac{r-1}{p}, \frac{r}{p} \right]$  then
8         | | | |  $Elementos_r = Elementos_r + 1$ ;
9     | | | return  $Elementos_r$  ;
```

---



---

**Algorithm 3.7.4:** Fase 2: Frecuencia de elementos

---

**Data:**  $Elementos_r, M$

**Result:**  $Frec_r$

```

1 Inicializa  $|Src|$ :  $|Src| = NumFil(M)$ ;
2 foreach  $x_i \in Src$  do
3   | foreach  $r \in p$  do
4     | |  $Frec_r = \frac{Elementos_r}{|Src|}$ ;
5   | return  $Frec_r$  ;
```

---

<sup>8</sup>Utilizando las notaciones de la Sección 3.6. Fase 2: Cálculo de la Frecuencia de Scores

- **FASE 3: Cálculo del Valor de la Información** El cálculo del Valor de la Información se calcula de acuerdo a la Ecuación 3.11 que consiste en utilizar la Frecuencia de elementos que se halló en la fase anterior y aplicar la Ecuación del Valor de la Información Modificada 3.3 (Página 45)

---

**Algorithm 3.7.5:** Fase 3

---

**Data:**  $\overline{Frec}_r, p, M$   
**Result:**  $IV$ : Matriz del Valor de la Información

- 1 Inicializa  $|Mtd|$ :  $|Mtd| = NumCol(M)$ ;
- 2 Inicializa  $|Src|$ :  $|Src| = NumFil(M)$ ;
- 3 Inicializa  $Elementos_r = 0, \forall r \in p$ ;
- 4 Inicializa  $IV = Dim[|Mtd|, |Src|]$ ;
- 5 **foreach**  $k \in Mtd$  **do**
- 6     **foreach**  $x_i \in Src$  **do**
- 7         **foreach**  $r \in p$  **do**
- 8             **if**  $M(k, i) \in \left[ \frac{r-1}{p}, \frac{r}{p} \right]$  **then**
- 9                  $Frec = \overline{Frec}_r$ ;
- 10                 **Break** ;
- 11              $IV(k, i) = -\log_2(Frec) + \left( \frac{1}{Frec} \right) \log_2 |Mtd|$ ;
- 12 **return**  $IV$  ;

---

Después del cálculo de la Matriz del Valor de la Información ( $IV$ ) se le deberá linealizar, para esto se aplica un algoritmo similar al presentado en la *Fase 1* con la diferencia que esta vez la Matriz de entrada en la Matriz del Valor de la Información  $IV$  en lugar de la Matriz de Scores  $M$ , por ello no se colocará el respectivo algoritmo.

Sin embargo, y como ya se había explicado en la *Fase 1*, pueden existir *Valores atípicos* cuyo resultado era igual a  $-1$  para evitar que estos valores causen algún conflicto a los resultados finales se realiza una actualización de la Matriz  $IV$  como sigue:

---

**Algorithm 3.7.6:** Fase 3: Actualización de  $IV$

---

**Data:**  $IV$ : Matriz del Valor de la Información,  $M$ : Matriz de Scores linealizada  
**Result:**  $IV$ : Matriz del Valor de la Información linealizada

- 1 Inicializa  $|Mtd|$ :  $|Mtd| = NumCol(IV)$ ;
- 2 Inicializa  $|Src|$ :  $|Src| = NumFil(IV)$ ;
- 3 **foreach**  $k \in Mtd$  **do**
- 4     **foreach**  $x_i \in Src$  **do**
- 5         **if**  $IV(k, i) = -1$  **then**
- 6              $IV(k, i) = \frac{1}{M(k, i)}$ ;
- 7 **return**  $IV$  ;

---

Con esta actualización hacemos que en la siguiente fase (Fase 4). Los resultados que presenten casos atípicos tomen un valor igual a uno luego del proceso, que para fines de este *Modelo de Fusión de Datos* es un valor neutro.



- **FASE 4: Integración de Scores y del valor de la información** Esta Fase es equivalente a la *Fase 4* del *Modelo de Fusión de Datos*. En esta etapa se integran, mediante multiplicación, los resultados de la Matriz de Scores linealizada  $M$  (Fase 1) con los resultados de la Matriz del Valor de la Información linealizada  $IV$  (Fase 3).

---

**Algorithm 3.7.7:** Fase 4

---

**Data:**  $M$ : Matriz de Scores linealizada,  $IV$ : Matriz del Valor de la Información

**Result:**  $\widetilde{M}$ : Matriz Integrada

```

1 Inicializa  $|Mtd|$ :  $|Mtd| = NumCol(M)$ ;
2 Inicializa  $|Src|$ :  $|Src| = NumFil(M)$ ;
3 foreach  $k \in Mtd$  do
4   |   foreach  $x_i \in Src$  do
5     |   |  $\widetilde{M}(k, i) = M(k, i) \cdot IV(k, i)$ ;
6 return  $\widetilde{M}$  ;
```

---

En este punto, el algoritmo puede modificar y en lugar de entregar un valor en escala entera, puede entregar un valor en escala logarítmica como se hacía referencia en el *Modelo de Fusión de Datos Integrado* (Página 53). Siendo sólo necesario un cambio menor en el algoritmo:

---

**Algorithm 3.7.8:** Fase 4: Escala logarítmica

---

**Data:**  $M$ : Matriz de Scores linealizada,  $IV$ : Matriz del Valor de la Información

**Result:**  $\log_{10} \widetilde{M}$ : Matriz Integrada en escala logarítmica

```

1 Inicializa  $|Mtd|$ :  $|Mtd| = NumCol(M)$ ;
2 Inicializa  $|Src|$ :  $|Src| = NumFil(M)$ ;
3 foreach  $k \in Mtd$  do
4   |   foreach  $x_i \in Src$  do
5     |   |  $\log_{10} \widetilde{M}(k, i) = \log_{10} M(k, i) + \log_{10} IV(k, i)$ ;
6 return  $\log_{10} \widetilde{M}$  ;
```

---

- FASE 5: Fusión de Scores** La última Fase del *Modelo de Fusión de Datos* es la Fusión de Scores que se realiza mediante la *Combinación Geométrica* presentada en la sección 3.4, y además se toma el concepto del *Factor de Credibilidad* de la sección 3.5 que es un dato externo ingresado por el usuario<sup>9</sup>. Para este fin, se realiza un algoritmo en tres etapas: (a) Agregar el Factor de Credibilidad a la ecuación, (b) Utilizar la Combinación Geométrica, (c) Sacar la raíz k-ésima.

---

**Algorithm 3.7.9:** Fase 5: Factor de Credibilidad

---

**Data:**  $\widetilde{M}$ : Matriz Integrada,  $FC_k$ : Factor de credibilidad para el k-ésimo método  $\forall k \in Mtd$   
**Result:**  $\widetilde{M}$ : Matriz Integrada con Factor de Credibilidad

- 1 Inicializa  $|Mtd|$ :  $|Mtd| = NumCol(\widetilde{M})$ ;
- 2 Inicializa  $|Src|$ :  $|Src| = NumFil(\widetilde{M})$ ;
- 3 **foreach**  $k \in Mtd$  **do**
- 4     **foreach**  $x_i \in Src$  **do**
- 5          $\widetilde{M}(k, i) = \widetilde{M}(k, i)^{FC_k}$ ;
- 6 **return**  $\widetilde{M}$  ;

---



---

**Algorithm 3.7.10:** Fase 5: Combinación Geométrica

---

**Data:**  $\widetilde{M}$ : Matriz Integrada  
**Result:**  $M_F$ : Vector Fusionado

- 1 Inicializa  $|Mtd|$ :  $|Mtd| = NumCol(\widetilde{M})$ ;
- 2 Inicializa  $|Src|$ :  $|Src| = NumFil(\widetilde{M})$ ;
- 3 Inicializa  $M_F = Dim[|Src|] = 1$ ;
- 4 **foreach**  $x_i \in Src$  **do**
- 5     **foreach**  $k \in Mtd$  **do**
- 6          $M_F(i) = M_F(i) \cdot \widetilde{M}(k, i)$ ;
- 7 **return**  $M_F$  ;

---



---

**Algorithm 3.7.11:** Fase 5: Raíz k-ésima

---

**Data:**  $M_F$ : Vector Fusionado,  $|Mtd|$ : Número total de Métodos de Detección de Similitud  
**Result:**  $M_F$ : Vector Fusionado Final

- 1 Inicializa  $|Src|$ :  $|Src| = NumFil(\widetilde{M})$ ;
- 2 **foreach**  $x_i \in Src$  **do**
- 3      $M_F(i) = \sqrt{|Mtd|} M_F(i)$ ;
- 4 **return**  $M_F$  ;

---

<sup>9</sup>El Factor de Credibilidad servirá para calibrar el Modelo de Fusión de Datos para que responda mejor a un universo de Documentos con ciertas características. Se recomienda que la elección de este factor sea mediante juicio experto.

Al igual como ocurrió en la *Fase 4*, la *Fase 5* también permite que se apliquen escalas logarítmicas en las operaciones de cálculo. Esto ayuda a que las expresiones matemáticas sean menos complicadas para el computador que lo resuelve, agilizando de este modo el tiempo de respuesta del *Modelo de Fusión de Datos por Score*. Por ejemplo, se evita el cálculo de raíces k-ésimas y se les reemplaza por el producto de una fracción, y los productos por sumas.

---

**Algorithm 3.7.12:** Fase 5: Factor de Credibilidad logarítmica

---

**Data:**  $\widetilde{M}$ : Matriz Integrada,  $FC_k$ : Factor de credibilidad para el k-ésimo método  $\forall k \in Mtd$   
**Result:**  $\log_{10} \widetilde{M}$ : Matriz Integrada con Factor de Credibilidad

- 1 Inicializa  $|Mtd|$ :  $|Mtd| = NumCol(\widetilde{M})$ ;
- 2 Inicializa  $|Src|$ :  $|Src| = NumFil(\widetilde{M})$ ;
- 3 **foreach**  $k \in Mtd$  **do**
- 4     **foreach**  $x_i \in Src$  **do**
- 5          $\log_{10} \widetilde{M}(k, i) = FC_k \cdot \log_{10} \widetilde{M}(k, i)$ ;
- 6 **return**  $\log_{10} \widetilde{M}$  ;

---



---

**Algorithm 3.7.13:** Fase 5: Combinación Geométrica logarítmica

---

**Data:**  $\log_{10} \widetilde{M}$ : Matriz Integrada  
**Result:**  $\log_{10} M_F$ : Vector Fusionado

- 1 Inicializa  $|Mtd|$ :  $|Mtd| = NumCol(\log_{10} \widetilde{M})$ ;
- 2 Inicializa  $|Src|$ :  $|Src| = NumFil(\log_{10} \widetilde{M})$ ;
- 3 Inicializa  $\log_{10} M_F = Dim[|Src|] = 1$ ;
- 4 **foreach**  $x_i \in Src$  **do**
- 5     **foreach**  $k \in Mtd$  **do**
- 6          $\log_{10} M_F(i) = \log_{10} M_F(i) + \log_{10} \widetilde{M}(k, i)$ ;
- 7 **return**  $\log_{10} M_F$  ;

---



---

**Algorithm 3.7.14:** Fase 5: Raiz k-ésima logarítmica

---

**Data:**  $\log_{10} M_F$ : Vector Fusionado,  $|Mtd|$ : Número total de Métodos de Detección de Similitud  
**Result:**  $\log_{10} M_F$ : Vector Fusionado Final

- 1 Inicializa  $|Src|$ :  $|Src| = NumFil(\log_{10} \widetilde{M})$ ;
- 2 **foreach**  $x_i \in Src$  **do**
- 3      $\log_{10} M_F(i) = \frac{1}{|Mtd|} \cdot \log_{10} M_F(i)$ ;
- 4 **return**  $\log_{10} M_F$  ;

---

## Capítulo 4

# Experimentaciones

En este capítulo se detallan los pasos seguidos para las experimentaciones del *Modelo de Fusión de Datos por Score* propuesto en el Capítulo 3, el proceso de experimentación se construyó siguiendo un esquema que toma cierta similitud con el *proceso KDD*<sup>1</sup>, el cual se denominó *proceso de detección de plagio en documentos* o *proceso DPD* cuyo esquema se presenta en la Figura 4.1.

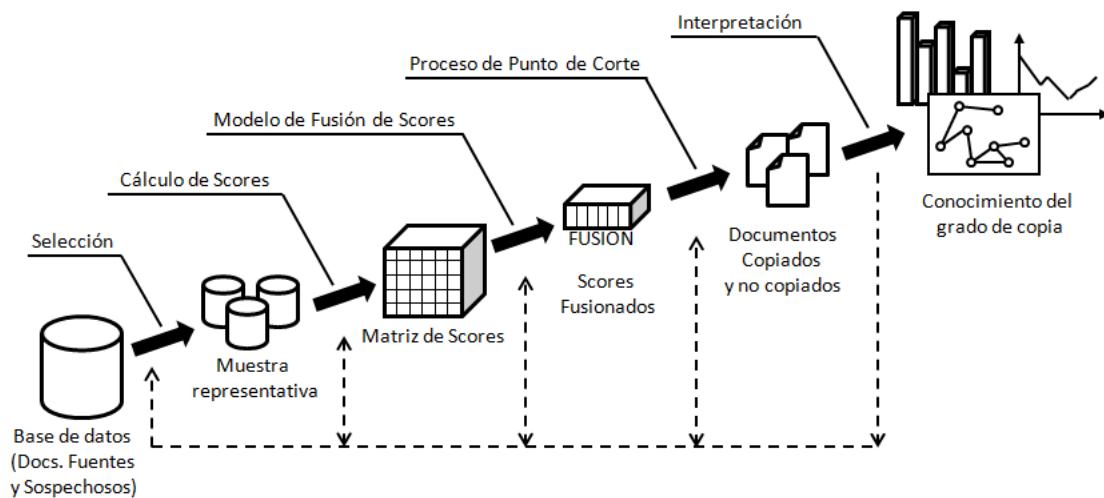


Figura 4.1: Proceso de Detección de Plagio de Documentos (Proceso DPD)

El proceso consta de cinco etapas: **(a) Selección.** A partir de una base de datos que contiene *Documentos Sospechosos* y *Documentos Fuentes* se extrae una muestra representativa del universo total de documentos; esto puede ser un *Documento Sospechoso* y muchos *Documentos Fuentes* o muchos *Sospechosos* y *Fuentes*. **(b) Cálculo de Scores.** Con el uso de diversos *Métodos de Detección de Similitud* con el

<sup>1</sup>Knowledge Discovery in Databases: Proceso no trivial para identificar patrones previamente desconocidos, válidos, nuevos, potencialmente útiles y comprensibles dentro de los datos

cual se evalúan *el(los) Documento(s) Sospechoso(s)* con los *Documentos Fuentes* y se construye una *Matriz de Scores* tal como se explicó en la Sección 3.2 del Capítulo 3. **(c) Modelo de Fusión de Datos.** Mediante un *Modelo de Fusión de Datos* se integran todos los Scores que se encuentran en la Matriz para obtener un *Vector de Scores Fusionados* (estos conceptos fueron explicados en la sección 3.2). **(d) Punto de Corte.** Luego de conocer el *Score Fusionado* se le puede ordenar descendientemente para conocer qué *Documentos Fuentes* tienen más probabilidad de ser el origen del texto del *Documento Sospechoso* sin embargo se requiere de un valor de corte  $c$  (Donde:  $c \in [0; 1]$ ) que ayude a discriminar exactamente de qué Documentos Fuentes provino el Documento Sospechoso. Con este proceso se conseguirá conocer el número de Documentos copiados y el número de Documentos no copiados que existen en la *Muestra representativa* de la primera etapa. **(e) Interpretación de resultados.** Después de conocer para cada *Documento Sospechoso* sus *Fuentes de copia* se pueden realizar análisis e interpretaciones de los resultados y obtener datos estadísticos para, por ejemplo, conocer la tendencia de copia, el análisis de colusión, entre otros.

Para las primera pruebas del *Modelo de Fusión de Datos por Score* se utilizará una *Base de Datos* con valores supervisados, es decir se conocerán a priori los *Documentos Fuentes* para cada *Documento Sospechoso* y se seguirán todas las etapas del *proceso DPD* a excepción de la última (Interpretación) que será cambiada por un *Análisis estadístico* que permitirá conocer la precisión y eficiencia del *Modelo* de acuerdo a los indicadores del Apéndice B. Detallando cada una de las etapas del *proceso DPD* se tienen cinco sub-secciones que se presentan a continuación.

## 4.1. Selección de datos

Para las experimentaciones se utilizó como fuente de información la Base de Datos **PAN plagiarism corpus 2010** (PAN-PC-10)<sup>2</sup> que es una base de libre descarga administrada por un grupo de académicos de varios países como la Universität Weimar (Alemania), Universidad Politécnica de Valencia (España), University of the Aegean (Grecia) y Bar-Ilan University (Israel).

El grupo académico organiza anualmente un concurso y posterior workshop llamado *PAN: Uncovering Plagiarism, Authorship and Social Software Misuse* el cual es un evento que convoca a investigadores del área de las Tecnologías de la Información que están trabajando en *Sistemas de Detección de Plagio*. Para el concurso, la PAN libera una *Base de Datos supervisada*, llamada PAN plagiarism evaluation corpus, que contiene más de 20000 Documentos Sospechosos y más de 20000 Documentos Fuentes.

---

<sup>2</sup><http://pan.webis.de>. Último acceso: 15-Sep-2010

#### 4.1.1. PAN plagiarism corpus

De acuerdo al documento oficial de la PAN [55], el *PAN plagiarism corpus 2010* está formado a partir de más de 22000 libros electrónicos de libre acceso<sup>3</sup> los cuales se encuentran en un archivo de 1.7GB que posee las siguientes características:

- Contiene 20611 *Documentos Sospechosos* de copia
- Contiene 20612 *Documentos Fuentes* de copia
- Los documentos tienen tres tipos de longitudes: (a) Pequeños. De 1 a 10 páginas. (b) Medianos. De 11 a 100 páginas. (c) Grandes. De 100 a 1000 páginas. (Ver Figura 4.2 (a))
- La mitad de los *Documentos Sospechosos* presentan copia (Figura 4.2 (b))
- En los Documentos Sospechosos copiados, las palabras copiadas varían entre 50 a 5000 palabras.
- Del total de documentos (Documentos sospechosos y fuentes) 37100 documentos están en idioma *Inglés* y la otra parte entre *Español* y *Alemán*. (Figura 4.2 (c))

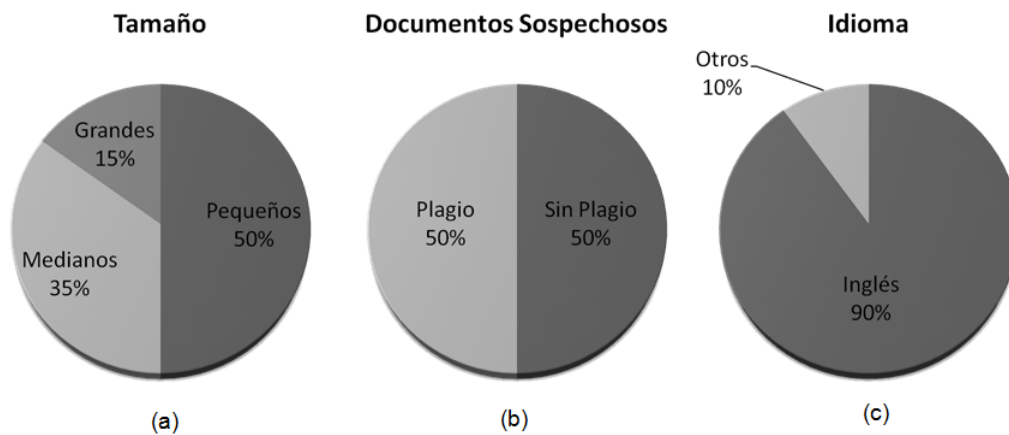


Figura 4.2: Distribución de los documentos del PAN plagiarism corpus 2010

#### 4.1.2. Lógica de selección de datos

Con los datos del PAN plagiarism corpus utilizado como Base de Datos supervisada, se realiza un histograma para conocer el número máximo de *Documentos Fuentes* con los que cuenta un *Documento Sospechoso*, se ordenan descendientemente y se enumeran los Documentos Sospechosos del 1 al n, quedando el gráfico de la Figura 4.3

<sup>3</sup>La fuente de estos libros es el Proyecto Gutenberg ([www.gutenberg.org](http://www.gutenberg.org)), el cual es un repositorio de libros electrónicos gratuitos que posee unos 30000 ejemplares.

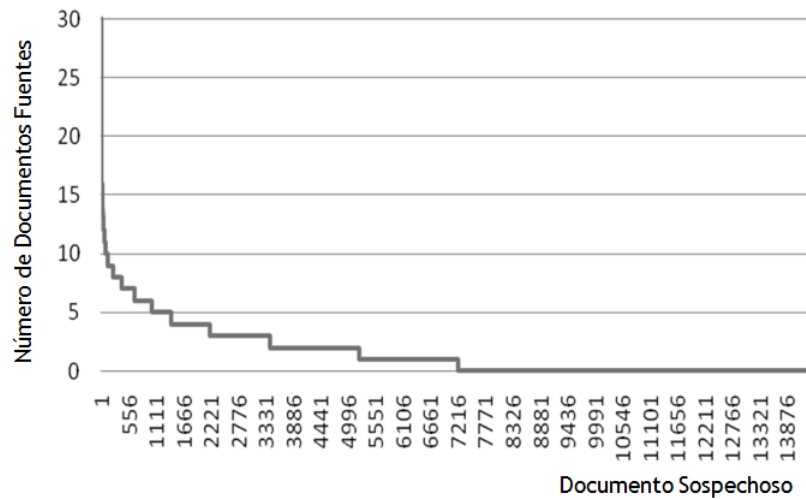


Figura 4.3: Histograma de los Documentos Fuentes y Sospechosos del PAN corpus 2010

Se decide tomar una muestra representativa que conste de 80 *Documentos Sospechosos*, 500 *Documentos Fuentes* y se realiza el filtro que sólo sean Documentos en idioma inglés. Quedando el histograma de la Figura 4.4, que se observa es una muestra representativa bastante fiel a la data original.

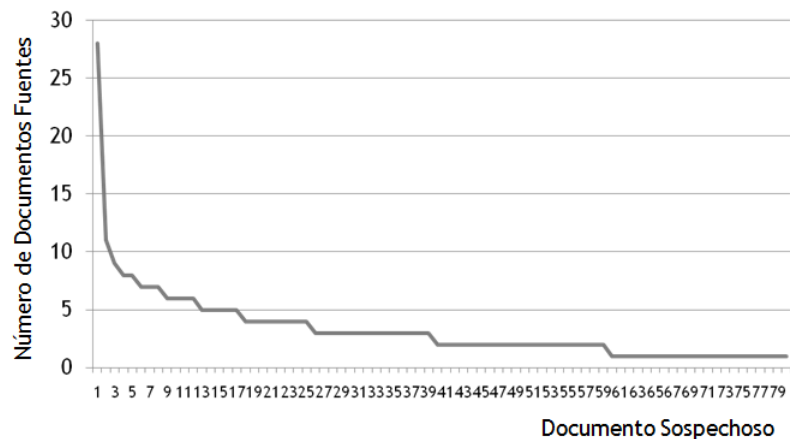


Figura 4.4: Histograma de los Documentos Fuentes y Sospechosos seleccionados

Además, los 80 *Documentos Sospechosos* elegidos deben contar con al menos un *Documento Fuente*, lo que se encuentra expresado en el cuadro 4.1; se eligieron 500 *Documentos Fuentes*. La lectura de este cuadro es, por ejemplo, para la primera línea: Hay 21 *Documentos Sospechosos* que poseen 1 *Documento Fuente*. Para la cuarta línea será: Hay 8 *Documentos Sospechosos* que poseen 4 *Documentos Fuentes*.

La lista de los 80 documentos elegidos del PAN Corpus 2010 se encuentran en el Apéndice D.

Número de Documentos Fuentes	Número de Documentos Sospechosos
1	21
2	20
3	14
4	8
5	5
6	4
7	3
8	2
9	1
11	1
28	1
<b>Total</b>	80

Cuadro 4.1: Documentos Sospechosos y el Número de Documentos Fuentes que contienen

## 4.2. Cálculo de Scores

Para realizar el *Cálculo de Scores* se utilizaron cuatro familias de *Métodos de Detección de Similitud entre Documentos* los cuales se subdividieron en tres tipos, quedando en total doce distintos *Métodos de Detección de Similitud* con los cuales se contrastó el *Modelo de Fusión de Datos*. Las familias de Métodos utilizados fueron: **(a) SimParalelo**. Es un método de detección de similitud de datos que analiza por *n*-gramas la similitud entre *Documento Sospechoso* y *Fuentes*, se llama *paralelo* porque puede procesar simultáneamente muchos grupos de *n*-gramas de los *Documentos Sospechoso* y *Fuentes*. **(b) SimTFIDE**. Tiene un funcionamiento similar al *SimParalelo* respecto a la comparación por *n*-gramas con la diferencia que la comparación no es exhaustiva sino que analiza sólo los *n*-gramas que contengan palabras con el mayor TF-IDF<sup>4</sup>. **(c) SimAlfabético**. También es un *Método de Similitud de Documentos* desarrollado por Oberreuter [53], al igual que los dos anteriores, y se encarga de comparar la similitud entre *n*-gramas con la particularidad que realiza comparaciones por orden alfabético. **(d) Diff**. Es el comparador de Similitud entred documentos básico que se encarga de indicar las diferencias entre dos archivos de texto.

### 4.2.1. SimParalelo

El *SimParalelo* como ya se describió es un método de detección de similitud que se encarga de comparar *n*-gramas entre el *Documento Fuente* y los *Documento Sospechosos*. La comparación *SimParalelo* utiliza los conceptos de *Ventanas*, *n*-gramas y *números de comparación*. Las ventanas son divisiones que se hacen sobre el documento completo, estas ventanas contienen un determinado grupo de palabras, que por construcción siempre debe contener 2 o más palabras y no ser del tamaño de todo el documento (Ver Figura

<sup>4</sup>Term Frequency – Inverse Document Frequency. Es un cuantificador utilizado en Text mining que permite reconocer qué *palabra* es más importante dentro de un texto



4.5). Los n-gramas son conjuntos de palabras: una palabra es un mono-grama, dos palabras un bi-grama, tres un tri-grama y así sucesivamente. El número de comparaciones es un contador interno del *Método* que se encarga de contar n-gramas iguales y consecutivos, cuando se encuentra un número de n-gramas igual al número de comparación entonces el Score aumenta.

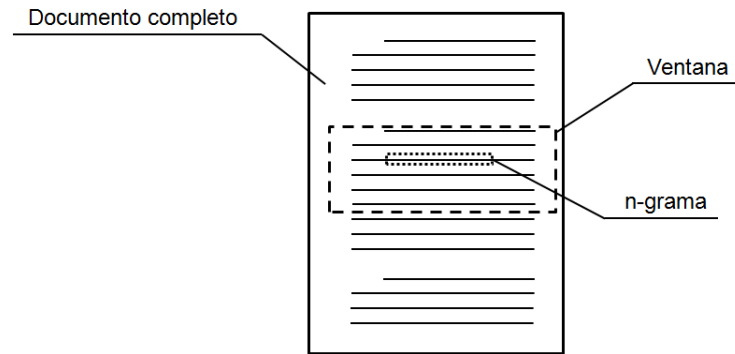


Figura 4.5: Detalle del funcionamiento del Método de Similitud

Este método permite un funcionamiento en paralelo de acuerdo a las capacidades del computador donde se le haga funcionar y dependerá del número de procesadores que tenga el computador. Mientras más procesadores, más funcionamientos en paralelo podrán ocurrir.

El SimParalelo presenta una buena eficiencia. Sin embargo, como procesa todos los n-gramas de cada ventana que pertenece a un documento su análisis es exhaustivo, y consecuentemente su tiempo de análisis es alto siendo dependiente del tamaño del documento.

$$Score = SimParalelo(n, m, k, c) \quad (4.1)$$

Donde

- $n$  : Número de n-gramas a analizar.
- $m$  : Número de palabras que pertenece en a la ventana.
- $k$  : Cantidad de comparaciones necesarias para aumentar el Score.
- $c$  : Número de núcleos (core) donde realizará el paralelismo.

Las variaciones elegidas para método se presentan en el siguiente cuadro:

#### 4.2.2. SimTFIDF

El funcionamiento del método de Similitud por TF-IDF toma una división del documento como lo hace el *SimParalelo* (ver Figura 4.5), pero agrega un análisis adicional para encontrar las palabras más representativas, o de más valor, en el texto. Las palabras se ordenarán ,de acuerdo al valor de su TF-IDF,

Nombre	Parámetros			
	n	m	k	c
SimPalarell0	3	5	3	16
SimPalarell1	2	6	3	16
SimPalarell2	4	8	3	16

Cuadro 4.2: Variaciones del SimParalell

desde la más representativa hasta la menos representativa entonces entrarán tres nuevas variables: las p-palabras más representativas, una *ventana* de q-palabras representativas y las r-palabras seleccionadas de la ventana q.

Esto es, si se tienen  $\tau$  palabras en el *Documento* se elegirá un sub-grupo de p-palabras (Donde  $p \leq \tau$ ). Luego, se aplicará una ventana de q-palabras sobre el sub-grupo (Donde  $q \leq p$ ) y de dicha ventana se elegirán sólo r-palabras (Donde  $r \leq q$ ).

El procesamiento con esta nueva lógica es mucho más rápido que con el *SimParalelo*, porque no evalúa exhaustivamente todas las palabras sino sólo aquellas con TF-IDF mayor; pero a cambio, pierde eficiencia.

$$Score = SimPTFIDF(n, m, k, p, q, r) \quad (4.2)$$

Donde

- $n$  : Número de n-gramas a analizar.
- $m$  : Número de palabras que pertenece en a la ventana.
- $k$  : Cantidad de comparaciones necesarias para aumentar el Score.
- $p$  : Cantidad de palabras con mejor TF-IDF del Documento.
- $q$  : Ventana de palabras aplicado a p.
- $r$  : Número de palabras elegidas en la ventana q.

Las variaciones elegidas para método son:

Nombre	Parámetros					
	n	m	k	p	q	r
SimTFIDF0	3	5	3	150	7	2
SimTFIDF1	2	6	3	150	7	2
SimTFIDF2	4	8	3	150	7	2

Cuadro 4.3: Variaciones del SimTFIDF

### 4.2.3. SimAlfabético

Este último método es bastante similar al *SimTFIDF*, posee la misma lógica de variables. La diferencia radica en que este método no ordena las palabras por TF-IDF sino por orden alfabético, siendo el primer carácter la letra a.

Sean  $\tau$  las palabras del *Documento* que se ordenan alfabéticamente y se elige un sub-grupo de  $p$ -palabras, se aplica una ventana de  $q$ -palabras al sub-grupo y de la mencionada ventana se extran  $r$ -palabras.

El tiempo de procesamiento es similar al *SimTFIDF*, por lo tanto mucho más rápido que el *SimParalelo*. Y los resultados son más precisos que el *SimTFIDF*, aunque nunca mejores que el *SimParalelo*

$$Score = SimAlfabetico(n, m, k, p, q, r) \quad (4.3)$$

Donde

- $n$  : Número de  $n$ -gramas a analizar.
- $m$  : Número de palabras que pertenece en a la ventana.
- $k$  : Cantidad de comparaciones necesarias para aumentar el Score.
- $p$  : Cantidad de palabras de mejor orden alfabético (iniciando en la letra a) del Documento.
- $q$  : Ventana de palabras aplicado a  $p$ .
- $r$  : Número de palabras elegidas en la ventana  $q$ .

Las variaciones elegidas para método son:

Nombre	Parámetros					
	n	m	k	p	q	r
SimALPH0	3	5	3	250	18	5
SimALPH1	2	6	3	250	18	5
SimALPH2	4	8	3	250	18	5

Cuadro 4.4: Variaciones del SimAlfabético

### 4.2.4. Diff

El último método de detección de similitud es bastante básico y no detecta con gran exactitud, sino que falla en muchos de los casos. Por dicha razón se le eligió porque más adelante se reconocerá que este método tiene un bajo *Factor de Credibilidad* (Ver Sección 3.5).

Este Sistema funciona reconociendo “similitud” entre líneas de dos documentos y responde si la línea del *Documento A* está en la misma posición que en el *Documento B*, se movió a otra zona, se eliminó o se le agregó alguna nueva línea al Documento B que no se encuentra en A.

El procesamiento de este sistema es bastante rápido, llega a superar a los tres métodos de detección de similitud mencionados anteriormente. Pero, como ya se había adelantado, su nivel de reconocimiento de similitud es bastante bajo

$$Score = Diff(add, mov, del) \quad (4.4)$$

Donde

- add* : Puntaje cuando una línea fue agregada al segundo Documento.
- mov* : Puntaje cuando una línea fue movida de lugar en el segundo Documento.
- del* : Puntaje cuando una línea fue eliminada del segundo Documento.

Cuando la línea se mantiene en la misma posición en el primer y segundo documento, el puntaje que se le suma es cero.

Los valores que ayudarán a variar los resultados del Diff elegidos se presetan a continuación

Nombre	Parámetros		
	add	mov	del
Diff0	-1	10	-1
Diff1	-10	0	-10
Diff2	-5	0	-10

Cuadro 4.5: Variaciones del Diff

### 4.3. Modelo de Fusión de Datos

Este paso consiste en aplicar el *Modelo de Fusión de Datos por Score* (Ver Sección 3.6) y con los datos de los dos pasos anteriores. Además, para cada método, se incluye el *Factor de Credibilidad* el cual se determinó mediante *juicio experto* por conocimiento previo de la eficiencia de los *Métodos de Detección de Similitud*; quedando la siguiente tabla de *Factor de Credibilidad*:

Nombre	Factor de Credibilidad
SimParalell0	0.7
SimParalell1	0.7
SimParalell2	0.7
SimTFIDF0	0.8
SimTFIDF1	0.8
SimTFIDF2	0.8
SimALPH0	0.9
SimALPH1	0.9
SimALPH2	0.9
Diff0	0.2
Diff1	0.2
Diff2	0.2

Cuadro 4.6: Factor de Credibilidad para los Método de Similitud de Documentos

Como se observa, se le tiene más confianza a los resultados del *Método* de similitud por criterio alfabético (SimALPH) porque en pruebas externas a esta tesis realizadas por Oberreuter [53] se concluyó que el método *SimAlfabético* detecta con mejores resultados la similitud entre Documentos que los métodos *SimParalell* y *SimTFIDF*. No se le otorgó una credibilidad de 1 (100%) porque se quería experimentar considerando que ningún *Método de Detección de Similitud* es perfecto.

En contraste a lo anterior, al *método Diff* se le asignó el “peor” *Factor de Credibilidad* porque en pruebas realizadas previamente, y como se indicó en la sub-sección 4.2.4 de este capítulo, no es un método eficiente y su reconocimiento de similitud es muy bajo. A pesar de ser un método de tan poca calidad, fue elegido para la etapa de *experimentación* porque introduce “ruido” al usarlo y con ello se puede verificar el rol que cumple el *Factor de Credibilidad* dentro del *Modelo de Fusión de Datos* propuesto en esta tesis.

A los otros *Métodos de Detección de Similitud* se les asignó factores de credibilidad altos porque su eficiencia resultó siendo buena [53], pero no mejor que el *SimAlfabético*; los valores elegidos variaron en rango de 0,1 a pesar que debería ser mucho menor (en el orden de las centésimas).

## 4.4. Punto de Corte

Normalmente esta etapa consiste en elegir un valor de corte  $c$  que se encuentre entre cero y uno ( $c \in [0; 1]$ ), la elección del valor depender estrictamente del juicio experto del usuario del sistema. Este valor está relacionado con el tipo de Documentos a analizar por lo cual se considera una variable externa.

Como se cuenta con una *Base de Datos* supervisada, para las *Experimentaciones*, se opta por precalcular los indicadores de eficiencia de los *Sistemas de Detección de Copia* (PRECISION, RECALL, F-MEASURE) y utilizar un *Algoritmo Greedy* para hacer un análisis exhaustivo en un gran rango de probables valores para  $c$ , cuyo objetivo será encontrar el mejor corte  $c$  para obtener los mejores indicadores.

---

### Algorithm 4.4.1: Algoritmo Greedy para hallar el punto de Corte

---

**Data:**  $M_F$ : Vector Fusionado,  $minC$ : Mínimo valor de corte,  $maxC$ : Máximo valor de corte,  $stepC$ :

Paso de corte

**Result:**  $c$ : Mejor valor de corte

```

1 Inicializa  $c = 0$ ;
2 Inicializa  $step = minC$ ;
3 while  $step \in [minC; maxC]$  do
4   if  $Mejor(Precision(M_F), Recall(M_F), F-measure(M_F))$  then
5      $c = step$ ;
6    $step = step + stepC$ ;
7 return  $c$ ;
```

---

## 4.5. Comparación de resultados

La última etapa del *proceso DPD* es la interpretación de resultados, y esta sección también debió llamarse de ese modo porque a partir, de los *Documentos Copiados* y *No Copiados* se pueden realizar informes de grados de copia mediante diversas herramientas estadísticas. Siendo una de ellas el *Análisis de Colusión* que contiene grafos dirigidos (Ver Sección 2.2.4. Página 14). Sin embargo, como se desea comprobar que el *Modelo de Fusión de Datos* es eficiente entonces se calcularán y analizarán los indicadores de eficiencia del *Modelo de Fusión de Scores* en **comparación** con otros *Sistemas de Fusión* clásicos como los propuestos por Fox y Shaw [63]; Montague y Aslam [49] y que fueron presentados en la Sección 2.4.3 (Página 36)

Se opta por utilizar tres *técnicas de linealización* distintas: Estándar, Sumatoria y Z-score. Y tres *técnicas de combinación*: CombSUM, CombMNZ y CombANZ. Que al combinarlas entre sí, se obtienen nueve *Modelos de Fusión de Scores* distintos:

Los algoritmos que los *Modelos de fusión* utilizan, fueron divididos en cuatro: tres algoritmos de linealización, uno para cada método, y un algoritmo de combinación, que de acuerdo a los datos de ingreso

Modelo de Fusión	Descripción
Estándar - CombSUM	Modelo que integra la linealización estándar con la Combinatoria por sumatoria.
Estándar - CombMNZ	Modelo que integra la linealización estándar con la Combinatoria por <i>Mean Non Zero</i> .
Estándar - CombANZ	Modelo que integra la linealización estándar con la Combinatoria por <i>Average Non Zero</i> .
Sumatoria - CombSUM	Modelo que integra la linealización por sumatoria de scores con la Combinatoria por sumatoria.
Sumatoria - CombMNZ	Modelo que integra la linealización por sumatoria de scores con la Combinatoria por <i>Mean Non Zero</i> .
Sumatoria - CombANZ	Modelo que integra la linealización por sumatoria de scores con la Combinatoria por <i>Average Non Zero</i> .
Z-Score - CombSUM	Modelo que integra la linealización por varianza de scores con la Combinatoria por sumatoria.
Z-Score - CombMNZ	Modelo que integra la linealización por varianza de scores con la Combinatoria por <i>Mean Non Zero</i> .
Z-Score - CombANZ	Modelo que integra la linealización por varianza de scores con la Combinatoria por <i>Average Non Zero</i> .

Cuadro 4.7: Modelos clásicos de fusión por Score

se puede comportar como cualquier método de combinación.

---

**Algorithm 4.5.1:** Método de linealización estándar

---

**Data:**  $M$ : Matriz de Scores

**Result:**  $M$ : Matriz de Scores normalizada

1 Inicializa  $|Mtd| = NumCol(M)$ ;  $|Src| = NumFil(M)$ ;

2 **foreach**  $k \in Mtd$  **do**

3      $max = -\infty$ ;  $min = +\infty$ ;

4     **foreach**  $x_i \in Src$  **do**

5         **if**  $M(k, i) > max$  **then**

6              $max = M(k, i)$ ;

7         **if**  $M(k, i) < min$  **then**

8              $min = M(k, i)$ ;

9     **if**  $max \neq min$  **then**

10         **foreach**  $x_i \in Src$  **do**

11              $M(k, i) = \frac{M(k, i) - min}{max - min}$ ;

12     **else**

13         **foreach**  $x_i \in Src$  **do**

14              $M(k, i) = 0$

15 **return**  $M$  ;

---

**Algorithm 4.5.2:** Método de linealización por sumatoria

**Data:**  $M$ : Matriz de Scores  
**Result:**  $M$ : Matriz de Scores normalizada

```

1 Inicializa  $|Mtd| = NumCol(M)$ ;  $|Src| = NumFil(M)$ ;
2 foreach  $k \in Mtd$  do
3    $max = -\infty$ ;  $sum = 0$ ;
4   foreach  $x_i \in Src$  do
5     if  $M(k, i) < min$  then
6        $min = M(k, i)$ ;
7        $sum = sum + M(k, i)$ 
8   if  $sum \neq min$  then
9     foreach  $x_i \in Src$  do
10       $M(k, i) = \frac{M(k, i) - min}{sum - min}$ ;
11   else
12     foreach  $x_i \in Src$  do
13       $M(k, i) = 0$ 
14 return  $M$  ;

```

**Algorithm 4.5.3:** Método de linealización Z-score

**Data:**  $M$ : Matriz de Scores  
**Result:**  $M$ : Matriz de Scores normalizada

```

1 Inicializa  $|Mtd| = NumCol(M)$ ;  $|Src| = NumFil(M)$ ;
2 foreach  $k \in Mtd$  do
3    $max = -\infty$ ;  $sum = 0$ ;
4   foreach  $x_i \in Src$  do
5      $\mu = Media(M, k)$  ;
6      $\sigma = DevStd(M, \mu, k)$  ;
7   if  $sigma \neq 0$  then
8     foreach  $x_i \in Src$  do
9        $M(k, i) = \frac{M(k, i) - \mu}{\sigma}$ ;
10   else
11     foreach  $x_i \in Src$  do
12        $M(k, i) = 0$ 
13 return  $M$  ;

```



Donde, deben existir las siguientes sub-funciones:

---

**Algorithm 4.5.4:** Media: Valor promedio

---

**Data:**  $M$ : Matriz de Scores,  $k$ : Método de similitud k-ésimo

**Result:**  $\mu$ : Media

```

1 Inicializa  $|Src| = NumFil(M)$ ;
2 Inicializa  $sum = 0$ ;
3 foreach  $x_i \in Src$  do
4    $sum = sum + M(k, i)$ ;
5  $\mu = \frac{sum}{|Src|}$ ;
6 return  $\mu$ ;

```

---



---

**Algorithm 4.5.5:** DevStd: Desviación estándar

---

**Data:**  $M$ : Matriz de Scores,  $\mu$ : Media,  $k$ : Método de similitud k-ésimo

**Result:**  $\sigma$ : Desviación estándar

```

1 Inicializa  $|Src| = NumFil(M)$ ;
2 Inicializa  $sum = 0$ ;
3 foreach  $x_i \in Src$  do
4    $sum = sum + ((M(k, i) - \mu)^2)$ ;
5  $\sigma = \sqrt{\frac{sum}{|Src|}}$ ;
6 return  $\sigma$ ;

```

---

Para los *métodos de Combinación*, como se indicó, se utilizará un algoritmo que los unifica. Y, a partir de los métodos presentados en la Sección 2.4.3 (Página 38), se tiene:

$$Sc_F(x_i) = |R(x_i)|^t \left( \sum_{k=1}^{|Mtd|} [Sc_k(x_i)] \right) \quad (4.5)$$

$$R(x_i) = |\{x_i | x_i \in S_k; \forall k \in Mtd\}| \quad (4.6)$$

Donde:

- $R(x_i)$  : Relevancia del documento  $x_i$ . Indica el número de métodos  $Mtd$  que lo colocan al  $x_i$  como relevante
- $t$  : Varía de acuerdo al tipo de método de combinación que se desea utilizar. Ver Cuadro 4.8

Método de Combinación	t
CombSUM	0
CombMNZ	1
CombANZ	-1

Cuadro 4.8: Valores del factor  $t$

---

**Algorithm 4.5.6:** Método de Combinación

---

**Data:**  $M$ : Matriz de Scores normalizada,  $t$ : Tipo de método de Combinación

**Result:**  $M_F$ : Vector Fusionado

- 1 Inicializa  $|Mtd| = NumCol(M)$ ;
  - 2 Inicializa  $|Src| = NumFil(M)$ ;
  - 3  $CombSUM = MetodoCombSUM(M)$ ;
  - 4  $R = Relevancia(M)$ ;
  - 5 **foreach**  $x_i \in Src$  **do**
  - 6      $M_F(i) = R^t \cdot CombSUM(i)$ ;
  - 7 **return**  $M_F$  ;
- 

---

**Algorithm 4.5.7:** MetodoCombSUM: Método de Combinación por Sumatoria

---

**Data:**  $M$ : Matriz de Scores normalizada

**Result:**  $CombSUM$ : Vector Fusionado por Sumatoria

- 1 Inicializa  $|Mtd| = NumCol(M)$ ;
  - 2 Inicializa  $|Src| = NumFil(M)$ ;
  - 3 Inicializa  $CombSUM = Dim [|Src|] = 0$ ;
  - 4 **foreach**  $k \in Mtd$  **do**
  - 5     **foreach**  $x_i \in Src$  **do**
  - 6          $CombSUM(i) = CombSUM(i) + M(k, i)$ ;
  - 7 **return**  $CombSUM$  ;
- 

---

**Algorithm 4.5.8:** Relevancia: Vector que guarda las relevancias de los documentos fuente  $Src$

---

**Data:**  $M$ : Matriz de Scores normalizada

**Result:**  $R$ : Vector de relevancia

- 1 Inicializa  $|Mtd| = NumCol(M)$ ;
  - 2 Inicializa  $|Src| = NumFil(M)$ ;
  - 3 Inicializa  $R = Dim [|Src|] = 0$ ;
  - 4 **foreach**  $k \in Mtd$  **do**
  - 5     **foreach**  $x_i \in Src$  **do**
  - 6         **if**  $M(k, i) \neq 0$  **then**
  - 7              $R = R + 1$ ;
  - 8 **return**  $R$  ;
-

## Capítulo 5

# Análisis de resultados

En este capítulo se revisarán y analizarán todos los resultados obtenidos luego del proceso de experimentación realizado (Capítulo 4). Primero se expondrá la evaluación para encontrar el *punto de corte* óptimo, después se presentarán y discutirán los resultados obtenido con el *Modelo de Fusión de Datos por Score*, que en los gráficos será llamado *MetaScore*<sup>1</sup>, respecto a diversos indicadores (Ver Apéndice B.1. A partir de los resultados, se hará un gráfico que muestre el número de fuentes para todos los documentos sospechoso para compararlo con los valores de la base de datos supervisada. Finalmente se realizará una comparación del *Método de Fusión de Datos propuesto* con otros métodos de fusión.

### 5.1. Resultados del Punto de Corte óptimo

Para el análisis del *Punto de Corte* óptimo, se utilizaron dos conceptos: los resultados supervisados de la *Base de Datos Seleccionada* para el uso de los indicadores de eficiencia (Ver Apéndice B) y el *algoritmo Greedy* del capítulo anterior (Sección 4.4, Algoritmo 4.4.1 Página 71); de este último se realizaron búsquedas en rangos: (a) **Primera búsqueda.** En la primera búsqueda se utilizó todo el espectro de valores para hallar el *punto de corte* se eligió el rango de 0 a 1, encontrándose que los mejores indicadores se encontraban muy cerca a 0. (b) **Segunda búsqueda.** Se analizó entre 0 y 0.004 con lo cual se encontró un máximo cerca a 0.003. (c) **Tercera búsqueda.** Se utilizó un rango más fino 0.00022 y 0.00042, esta vez el valor se encontraba entre 0.000316. (d) **Última búsqueda.** Por un tema de exactitud se quiso verificar el rango entre 0.0003122 y 0.0003922, lo cual sólo reforzó que el valor óptimo es el 0.000316.

En las Figuras 5.1 y 5.2 se muestra el comportamiento de los indicadores para la *Primera búsqueda*. Cabe resaltar que de los cuatro indicadores utilizados, el que más valor tuvo al momento de tomar la decisión del corte óptimo es el *F-measure* porque como se explica en el Apéndice B incluye una buena relación entre PRECISION y RECALL. En la Figura 5.1 se presenta el valor medio de los indicadores, mientras que en

---

<sup>1</sup>Esto no responde a ninguna otra razón que sólo ahorrar espacio en las leyendas de las imágenes

la Figura 5.2 se presenta la desviación estándar. Los valores exactos para cada uno de los 80 documentos seleccionados (Ver Sección 4.1), se encuentran detallados en en Apéndice D.

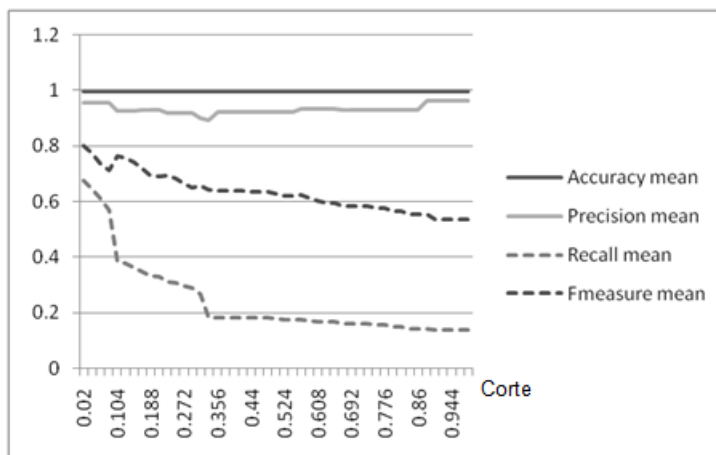


Figura 5.1: Media de los indicadores para el Punto de Corte óptimo

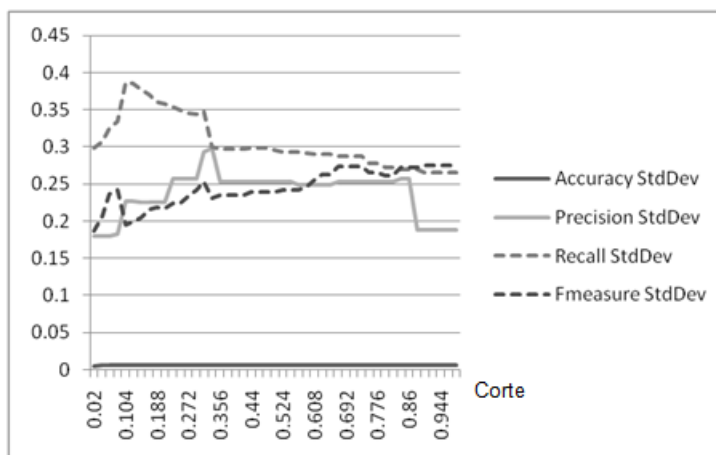


Figura 5.2: Desviación estándar de los indicadores para el Punto de Corte óptimo

Con este análisis se logra concluir que, para este tipo de documentos experimentales, el *valor de corte óptimo* está en  $3,6 \cdot 10^{-4}$ . Sin embargo, para buscar un mejor valor se elige como *Corte óptimo* el  $3,5 \cdot 10^{-4}$ .

Adicional a este paso, como estamos en la etapa de *Experimentaciones del Modelo de Fusión de Datos* se buscó verificar el mejor p-segmento, que al hacerlo funcionar entregue buenos indicadores.

## 5.2. Resultados del p-segmento óptimo

Similar a la sección anterior, para determinar el *mejor P-segmento* se utilizan los indicadores de eficiencia y el algoritmo Greedy, pero esta vez modificado para la variación del p-segmento en lugar del punto de corte. Consiguiéndose los gráficos de las Figuras 5.3 y 5.4.

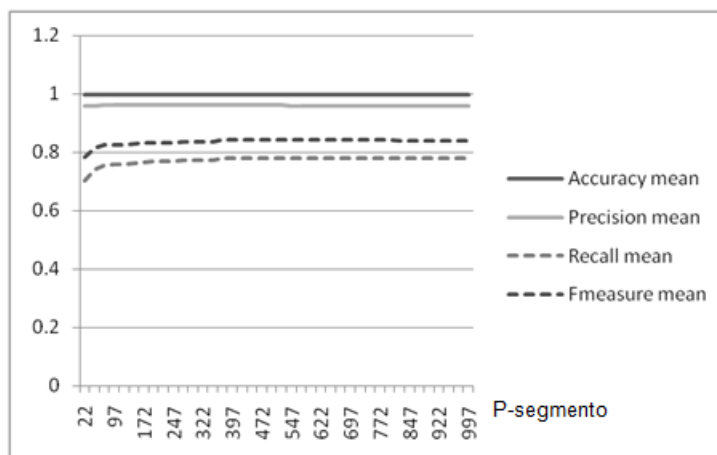


Figura 5.3: Media de los indicadores para el P-segmento óptimo

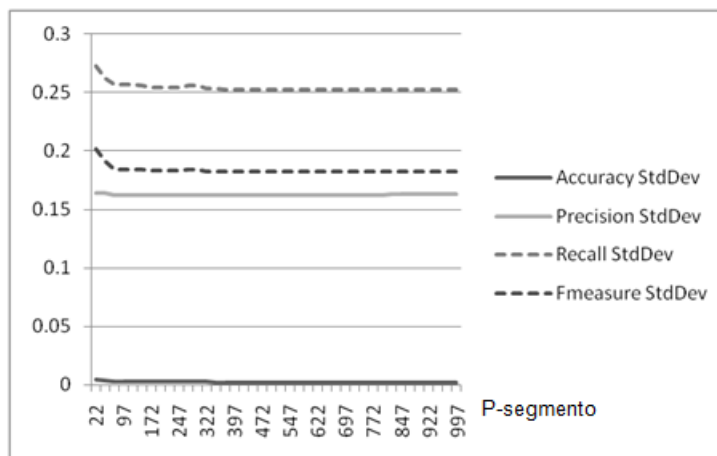


Figura 5.4: Desviación estándar de los indicadores para el P-segmento óptimo

Se consigue definir que el p-segmento óptimo se encuentra entre el rango de 475 y 525 segmentos. Por lo cual se decide utilizar un p-segmento de valor igual a 500. Cabe recalcar que los resultados numéricos de este análisis se encuentran en el Apéndice D.

### 5.3. Reconstrucción del número de Fuentes para cada Documento Sospechoso

Con los análisis previos se tienen dos valores óptimos que maximizan los resultados del *Modelo de Fusión de Datos* teniéndose así:

	Óptimos
Cut	$3,5 \cdot 10^{-4}$
P-segment	500

	Accuracy	Precision	Recall	F-measure
Mean	0.99815	0.96083333	0.78090233	0.84206821
StdDev	0.00218575	0.1625769	0.25281096	0.18246393

Cuadro 5.1: Indicadores para los valores óptimos

Con estos valores óptimos, se desea “reconstruir” la curva del histograma de la Figura 4.4 (Página 64), entonces se tendrá:

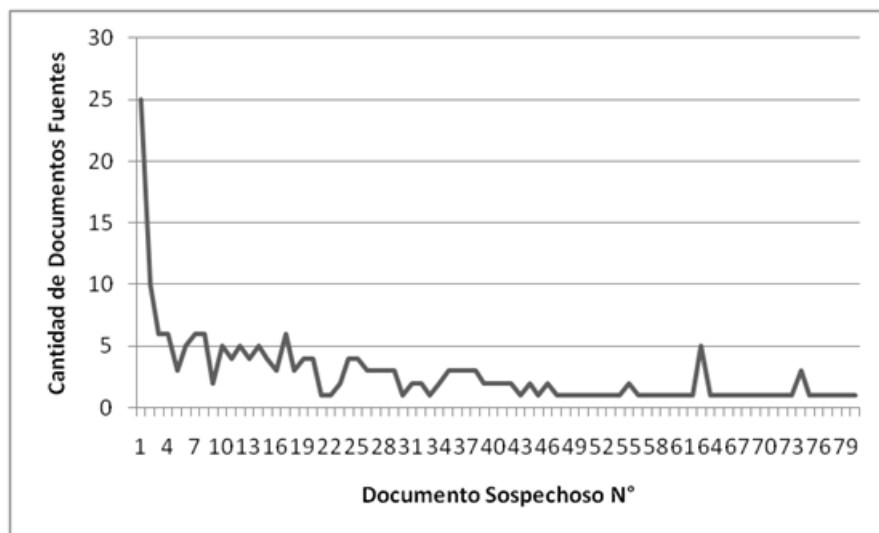


Figura 5.5: Reconstrucción del histograma del PAN Corpus 2010

Como se observa en la Figura, y como ya lo indicaba el Cuadro 5.1, existen casos donde la fusión de datos no fue correcta porque a pesar de lograr un ACCURACY muy cercano a uno, una precisión bastante buena y finalmente un F-MEASURE que supera el 84 % las desviaciones estándar también son altas en la PRECISION y el F-MEASURE. Esto se debe a que los *Métodos de detección de similitud* no funcionaron correctamente y no reconocieron si el *Documento Fuente* efectivamente era el origen de la información del *Documento Sospechoso*; y como dicen “*garbage in, garbage out*” entonces el *Método de Fusión de Datos*

por Score no pudo entregar buenos valores. La Figura 5.6 muestra la variación entre el resultados supervisado y el predicho.

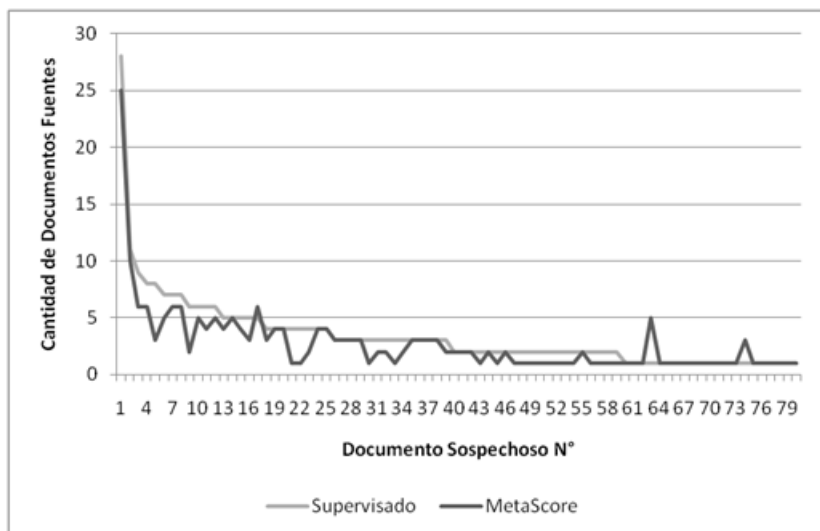


Figura 5.6: Comparación de histograma supervisado y predicho por el Método de Fusión de Scores

Sin embargo, se pueden mejorar los indicadores si se reconocen y filtran aquellos casos específicos donde los *Métodos de Detección de Similitud* fallan. Entonces: Quitando los 24 casos que causan conflicto en el reconocimiento de similitud; se tendrá una curva más suavizada (Ver Figura 5.7)

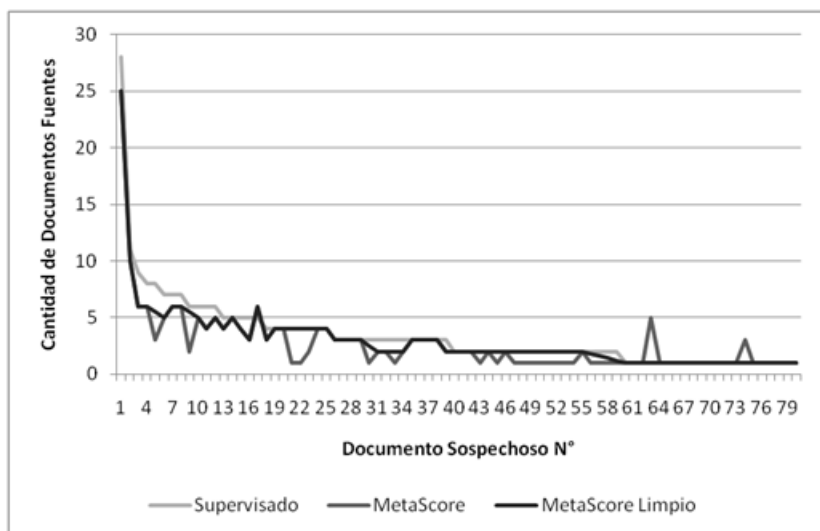


Figura 5.7: Modelo de Fusión de Datos con valores filtrados

Siendo sus resultados:

	<b>Óptimos</b>
Cut	$3,5 \cdot 10^{-4}$
P-segment	500

	<b>Accuracy</b>	<b>Precision</b>	<b>Recall</b>	<b>F-measure</b>
Mean	0.99884211	0.97076023	0.89644186	0.94365358
StdDev	0.00167321	0.1391733	0.17487494	0.08292523

Cuadro 5.2: Indicadores para los valores óptimos

Como se observa de la tabla 5.2 con el filtrado de aquellos documentos que causaban “ruido” a la Fusión de Datos se consiguió aumentar el F-MEASURE a una media de 94.3 % con una desviación estándar cercana al 8 %.

## 5.4. Métodos de Detección de Similitud

Para ahondar el motivo por el cual el *Modelo de Fusión de Datos por Score* falló en reconocer todos los documentos, se debe hacer un pre-análisis de los resultados promedios de los *Métodos de Detección de Similitud* como los del Cuadro 5.3

		<b>Accuracy</b>	<b>Precision</b>	<b>Recall</b>	<b>F-measure</b>
SimParalell	Mean	0.99853	0.90671	0.8913	0.88617
	StdDev	0.00204	0.20151	0.21102	0.164
SimTFIDF	Mean	0.92323	0.86742	0.78132	0.76745
	StdDev	0.26186	0.31843	0.26494	0.28549
SimALPH	Mean	0.89888	0.86955	0.82731	0.79084
	StdDev	0.29846	0.32212	0.23965	0.30061
Diff	Mean	0.38153	0.00531	0.56096	0.01719
	StdDev	0.37536	0.01009	0.48396	0.01986
<b>METASCORE</b>	<b>Mean</b>	<b>0.99815</b>	<b>0.96083</b>	<b>0.7809</b>	<b>0.84207</b>
	<b>StdDev</b>	<b>0.00219</b>	<b>0.16258</b>	<b>0.25281</b>	<b>0.18246</b>

Cuadro 5.3: Resultados de los Métodos de Similitud en la BD de prueba.

De todos estos métodos se realizó un gráfico comparativo para ver el reconocimiento de *Documentos copiados y no copiados* por tipo (Ver Figura 5.9).

Para la Figura 5.9, como todos los métodos reconocieron de manera diferente el número de *Documentos Fuentes* para cada *Documento Sospechoso*, se eligió la visualización en escala logarítmica. El método *Diff* tuvo el peor reconocimiento, asumiendo que los sospechosos tenían muchas fuentes (indicaba entre 119 a 500 fuentes por *Documento Sospechoso*). El Cuadro que muestra el número de fuentes para cada uno de los 80 Documentos sospechosos está en el Apéndice D.



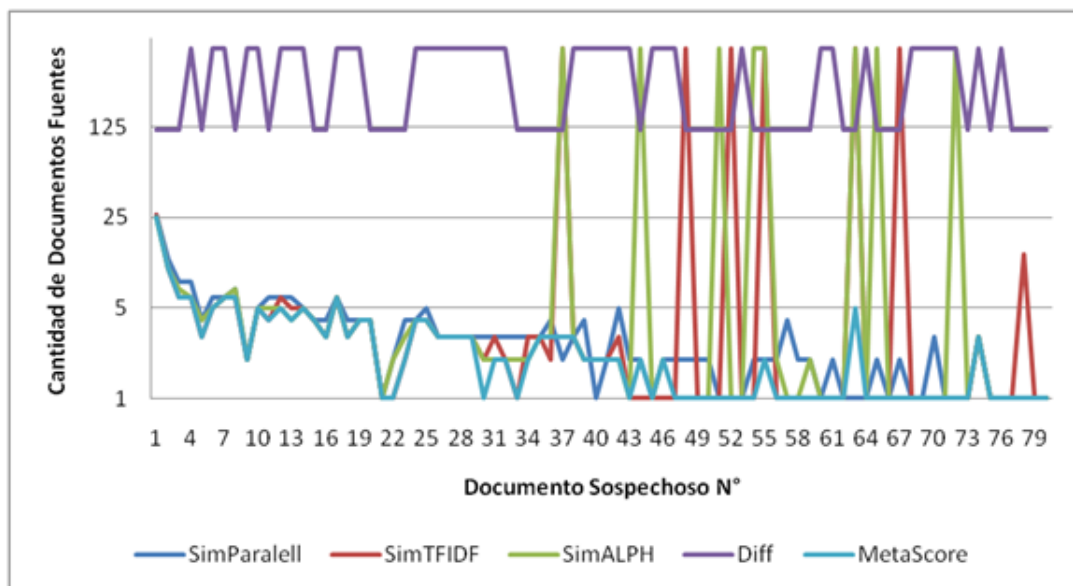


Figura 5.8: Documentos Fuentes vs. Documento Sospechoso por Método de similitud y de Fusión (escala logarítmica)

Si se quita la escala logarítmica del gráfico anterior, se obtiene la siguiente Figura, en la que se puede observar mejor como el *Modelo de Fusión de Datos propuesto* logró integrar adecuadamente a los diversos métodos de detección de similitud y a pesar del “ruido” introducido por el *Diff* el resultado final después de la Fusión no se alteró demasiado porque el *Factor de Credibilidad* cumplió un buen rol.

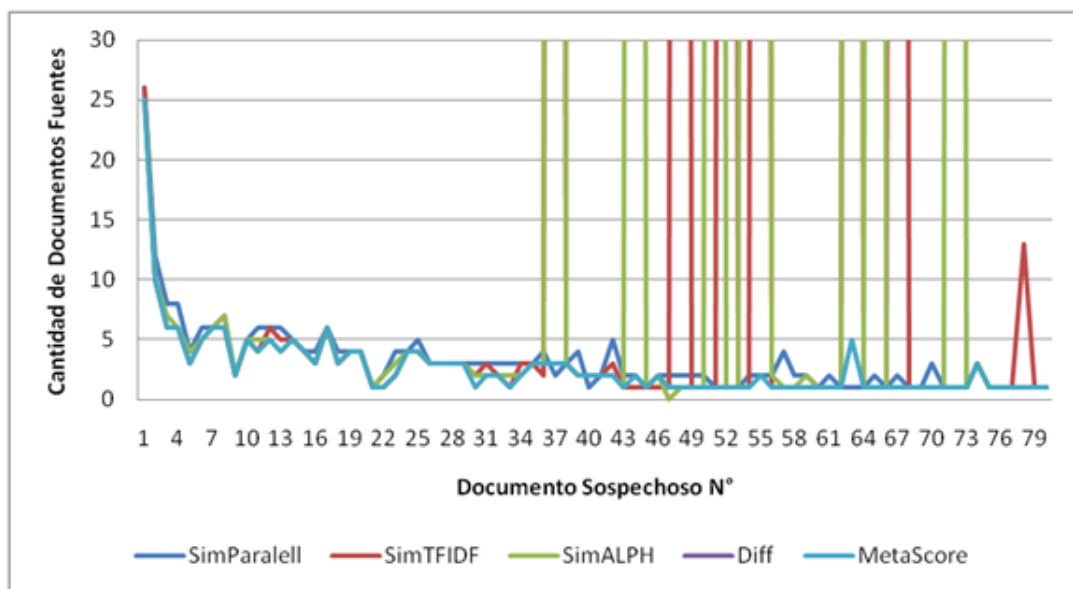


Figura 5.9: Documentos Fuentes vs. Documento Sospechoso por Método de similitud y de Fusión

## 5.5. Comparación con otros Modelos de Fusión de Datos

Como se explicó en la Sección 4.5 del Capítulo anterior se implementaron nueve distintos Modelos clásicos para fusión de Datos. A estos Modelos, al igual como se hizo con el *Modelo de Fusión de Datos propuesto*, se calculó un punto de corte óptimo que maximiza los indicadores de efectividad con lo que se obtuvieron los resultados del Cuadro 5.4.

Modelo de Fusión Clásico	Punto de corte
Estándar - CombSUM	2.350
Estándar - CombMNZ	8.490
Estándar - CombANZ	0.125
Sumatoria - CombSUM	0.190
Sumatoria - CombMNZ	0.350
Sumatoria - CombANZ	0.017
Z-Score - CombSUM	5.500
Z-Score - CombMNZ	3.500
Z-Score - CombANZ	0.690

Cuadro 5.4: Modelos clásicos de fusión por Score

Con los puntos de corte, se calcularon los indicadores para cada método (Ver Cuadro 5.5). De estos resultados se observa que el Modelo de Combinación Estándar-CombANZ respondió muy mal respecto a cualquier otro Modelo de Fusión, entregando una baja Precisión y por lo tanto un bajo F-MEASURE. Los demás Modelos tuvieron resultados buenos, hasta para algunos indicadores superaron la respuesta del *Modelo de Fusión de Datos propuesto*. En las Figuras 5.10, 5.11, 5.12, 5.13, 5.14 se analiza a detalle el comportamiento de los *Modelos de Fusión* para cada uno de los Indicadores.

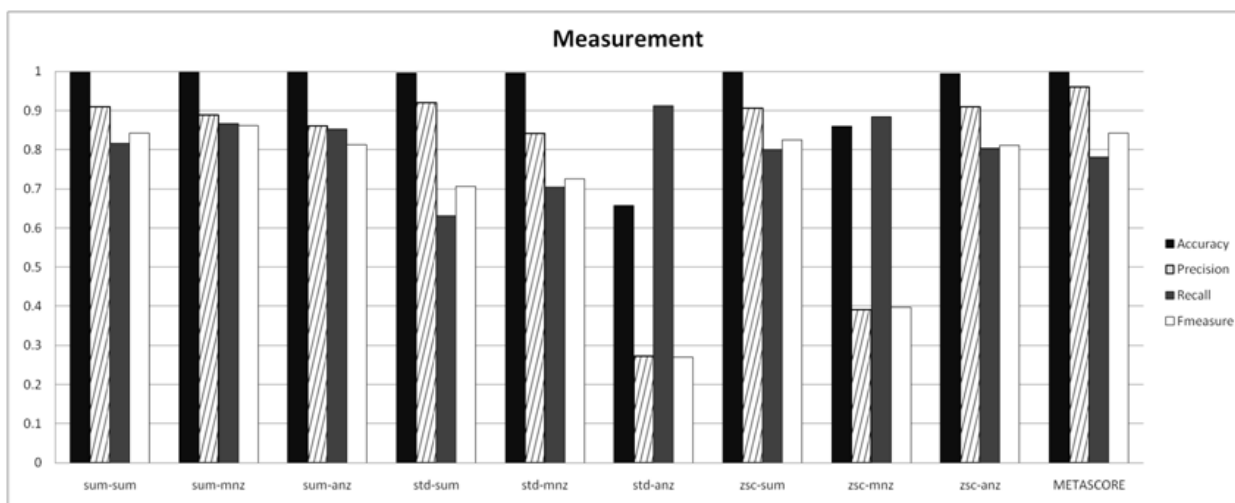


Figura 5.10: Comparación de Modelos de Fusión

En la Figura 5.10 se muestran todos los indicadores y todos los *Modelos de Fusión* a comparar; en el

		Accuracy	Precision	Recall	Fmeasure
Estándar CombSUM	Mean	0.9962	0.91979167	0.63139926	0.7059312
	StdDev	0.00567979	0.19783716	0.29756079	0.22358851
Estándar CombMNZ	Mean	0.996025	0.84199653	0.70487915	0.72471843
	StdDev	0.00633635	0.24524119	0.29108311	0.21681427
Estándar CombANZ	Mean	0.656975	0.27232809	0.91302985	0.27030442
	StdDev	0.36106537	0.41935346	0.16745653	0.39188169
Sumatoria CombSUM	Mean	0.99755	0.91041667	0.81697781	0.84257284
	StdDev	0.0045823	0.20530085	0.25230833	0.18820763
Sumatoria CombMNZ	Mean	0.9981	0.88854167	0.86780709	0.86254697
	StdDev	0.00248797	0.22819634	0.21286623	0.17443574
Sumatoria CombANZ	Mean	0.99715	0.86079545	0.85323323	0.81205236
	StdDev	0.00416863	0.24673573	0.21224317	0.20080148
Z-Score CombSUM	Mean	0.997425	0.90645833	0.80040584	0.82557277
	StdDev	0.00403972	0.21595225	0.25168346	0.19163746
Z-Score CombMNZ	Mean	0.859125	0.39084798	0.88386319	0.39669283
	StdDev	0.11680704	0.45611547	0.1892961	0.43702432
Z-Score CombANZ	Mean	0.994425	0.91051913	0.80492424	0.81045436
	StdDev	0.02696052	0.2147775	0.24622624	0.21778692
<b>METAScore</b>	<b>Mean</b>	<b>0.9981520</b>	<b>0.9608333</b>	<b>0.7809023</b>	<b>0.8420682</b>
	<b>StdDev</b>	<b>0.0021858</b>	<b>0.1625769</b>	<b>0.2528112</b>	<b>0.1824639</b>

Cuadro 5.5: Indicadores para los Modelos de Fusión Clásicos y para el MetaScore (Modelo propuesto)

caso de las imágenes se cambió el nombre del *Modelo de Fusión de Datos por Score* a *MetaScore* para fines prácticos de un nombre más breve. Se observa, como se indicó antes, que el *Modelo de Fusión Estándar* y *CombANZ* tuvo el peor desempeño siguiéndole el *modelo de Fusión Z-score* y *CombMNZ*, esto también se observa en la Figura 5.11 del indicador del ACCURACY.

La mayoría de *Modelos* obtuvo un buen ACCURACY y el *Modelos de Fusión de Datos propuesto* tampoco fue la excepción.

En la Figura 5.12 se tiene la comparación respecto al indicador de PRECISION, este indicador muestra que “De los  $n$  documentos que el Modelo dice que es copia, el  $m$  realmente lo es” (Donde  $m \leq n$ ). En este caso del *MetaScore* tuvo el mejor PRECISION (0,96) de todos los *Modelos* que se comparan, y en segundo lugar con 4 % menos de PRECISION (0,92) está el *Modelo de Fusión Estándar-CombSUM*. Esto indica que la confianza que se le puede tener al *MetaScore* es alta.

Con el indicador del RECALL (Figura 5.13), el *MetaScore* no tuvo el mejor resultado sin embargo su valor tampoco fue bajo (0,78). Dentro del contexto de la detección de plagio tener un alto RECALL indica que “De  $n$  documentos copiados que existían, el Modelo reconoció a  $m$  de ellos” (Donde:  $m \leq n$ ). El Modelo de Fusión con mejor RECALL fue el Estándar-CombANZ con 0,91; y los dos peores fueron el Estándar-CombSUM y el Estándar-CombMNZ con RECALLS de 0,63 y 0,70 respectivamente. Lo que indica un comportamiento bastante interesante de los *Modelos de Fusión clásicos* porque la linealización Estándar

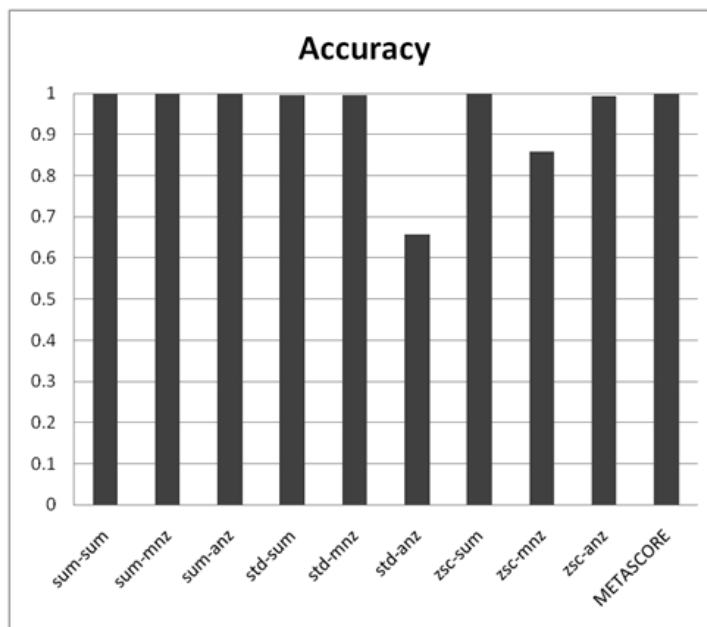


Figura 5.11: Comparación de Modelos por el ACCURACY

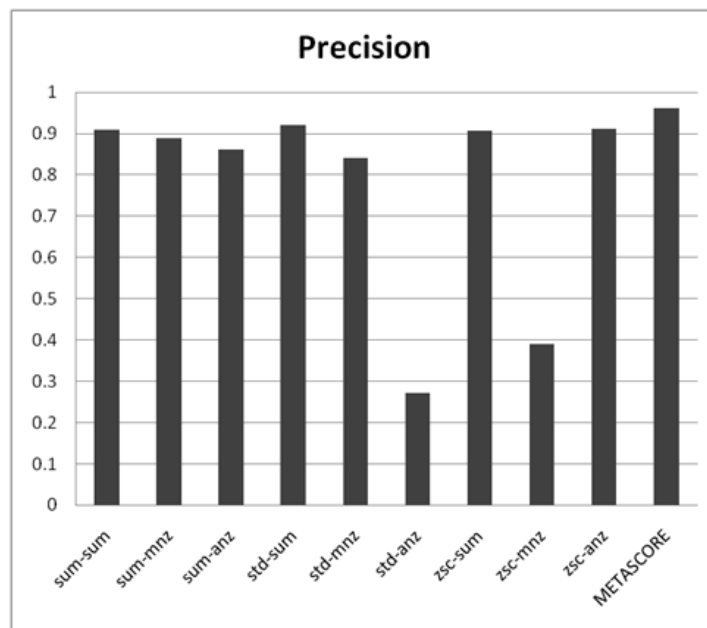


Figura 5.12: Comparación de Modelos por la PRECISION

entregó malos resultados en 2 de 3 casos, siendo el método de Combinación por Average Non-Zero (ANZ) que le hizo elevar su RECALL.

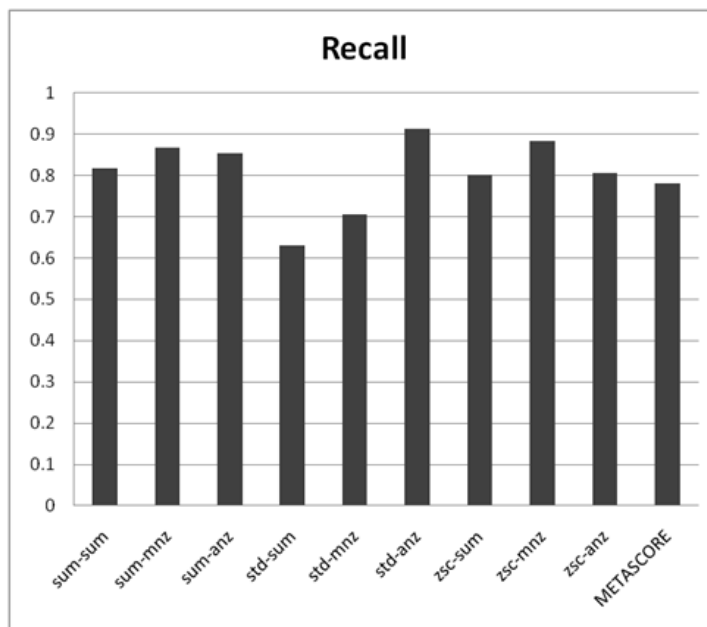


Figura 5.13: Comparación de Modelos por el RECALL

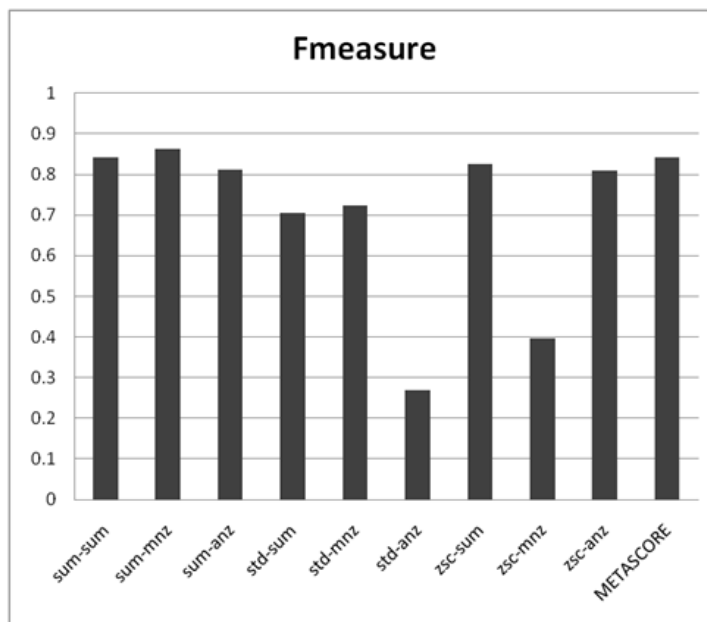


Figura 5.14: Comparación de Modelos por el F-MEASURE

Finalmente en la Figura 5.14 se muestra el indicador más representativo, el F-MEASURE, que es una función del PRECISION y el RECALL (Ver Apéndice B). De todos los *Modelos de Fusión de Datos* sólo dos de ellos *Sumatoria-CombSUM* y *Sumatoria-CombMNZ* superaron al desempeño del *MetaScore* pero por

diferencias muy bajas en el orden del 2 % y 0.05 % respectivamente.

En resumen, con la comparación de los *Modelos clásicos* y los respectivos indicadores de eficiencia se ha encontrado que el *Modelo de Fusión de Datos por Score propuesto* resulta eficiente. Además como lo demostró Yu Suzuki [67] los métodos que utilizan las linealizaciones clásicas (Ver Sección 2.4.3) no funcionan correctamente en todos los casos. Entonces se puede afirmar que si bien los *Modelos de Fusión Sumatoria-CombSUM* y *Sumatoria-CombMNZ* pueden presentar un mejor F-MEASURE para el caso supervisado del PAN Corpus 2010, en general el F-MEASURE del *MetaScore* será más confiable para casos genéricos.

## 5.6. Análisis de Colusión

Luego de comprobar que los resultados del *Modelo de Fusión de Datos por Score* entrega resultados fiables, se decidió aplicarlo sobre una tarea del ramo de *Tecnologías de la Información*<sup>2</sup> del *Magister en Gestión de Operaciones* dictado por el *Departamento de Ingeniería Industrial* de la *Universidad de Chile*.

Para esta prueba se eligieron las tareas 1 y 2 del semestre Otoño del 2010, donde habían: 35 Documentos en formatos (doc, docx y pdf) para la tarea 1, y 33 Documentos en formatos (doc, docx y pdf) para la tarea 2. Todos estos documentos se procesaron por un parser<sup>3</sup> y finalmente se consiguió: 32 Documentos para la tarea 1 y 30 para la tarea 2<sup>4</sup>.

La lógica de comparación varió un poco, porque esta vez todos los Documentos de las tareas podían ser *Documentos Sospechosos* y *Documentos Fuentes* al mismo tiempo (Ver Figura 5.15).

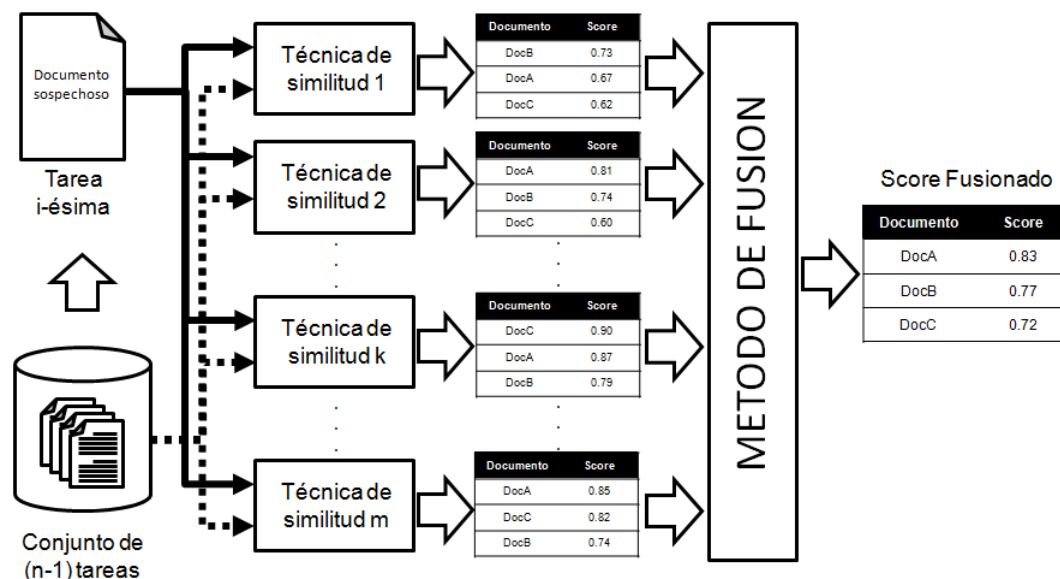


Figura 5.15: Lógica de análisis de copia entre Documentos para las tareas del IN72K

Del conjunto de  $n$  tareas del ramo se extrae una de ellas (Tarea  $i$ -ésima), esta tarea será considerada como *Documento Sospechoso* y las  $(n - 1)$  tareas restantes serán los *Documentos Fuentes* entonces se les procesa por cada uno de los *Métodos de Detección de Similitud*, se les fusiona y finalmente se obtiene un resultado del nivel de Similitud entre la Tarea  $i$  y las  $n - 1$  tareas restantes. Este proceso se repite  $n$  veces, uno por cada documento entregado como tarea. Luego se evalúan con el punto de corte para determinar las tareas que presentan similitudes con las otras. Por juicio experto se eligió un punto de corte igual a 0.3.

<sup>2</sup>El código del ramo en el año 2010 (cuando se desarrolló esta tesis) es IN72K y era dictado por el Profesor Sebastián Ríos

<sup>3</sup>Sistema que se encarga de extraer el texto de documentos que originalmente se encuentran en distintos formatos y entregar archivos en formato de texto (.txt)

<sup>4</sup>El número de documentos es menor porque algunos archivos no fueron reconocidos ni procesados adecuadamente por el parser

Con los resultados obtenidos por cada una de las  $n$  tareas, después de analizarlas por el punto de corte y por la ayuda del software libre Pajek, se puede representar en un grafo la relación de colusión (Ver Sección 2.2.4) entre los autores de las tareas del IN72K.

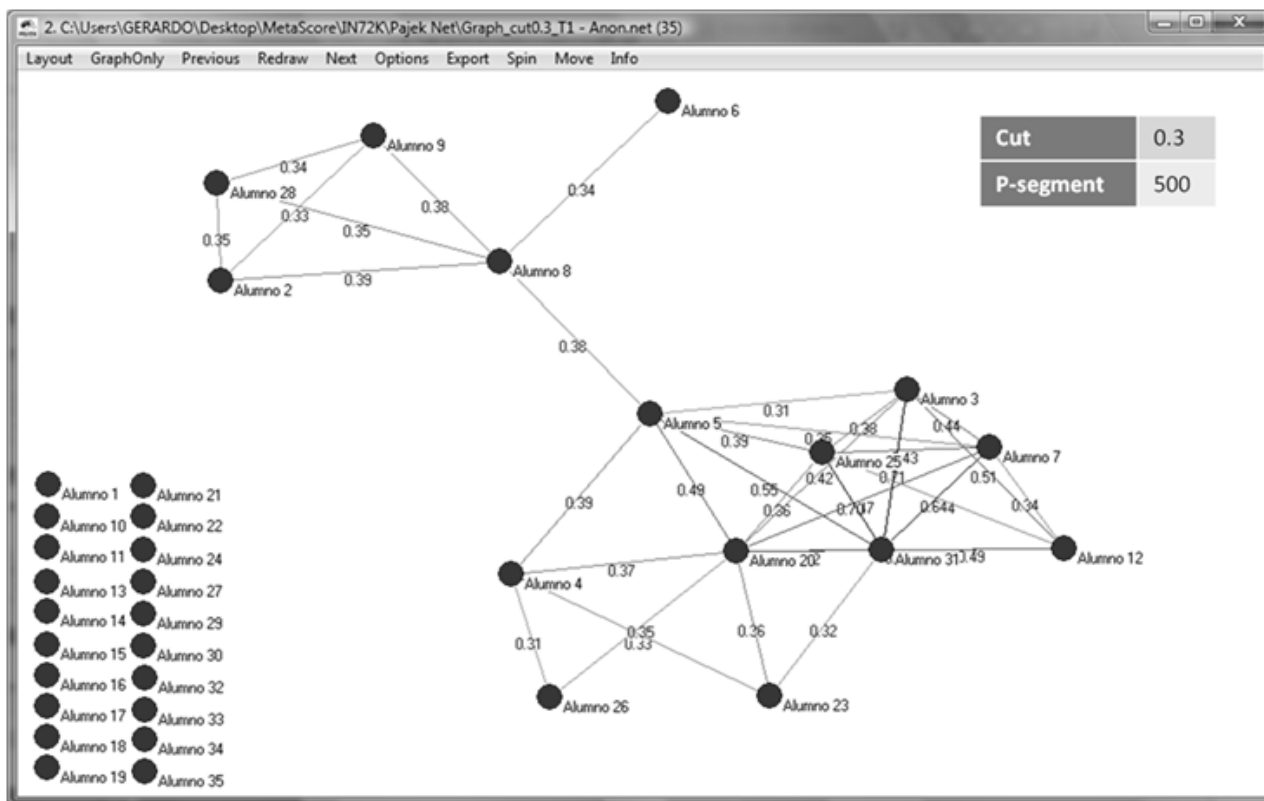


Figura 5.16: Análisis de Colusión para la tarea 1 del IN72K

Además, como se muestra en las Figuras 5.16 y 5.17, se consideró un valor de p-segmentos igual a 500. Esto debido a los análisis previos del p-segmento óptimo.

Para ambas tareas, los nombres se ocultaron con el objetivo de resguardar la identidad de los alumnos porque no se tiene permiso explícito de ellos para publicar esta información. En la tarea 1 (Figura 5.16) el grado de similitud era bastante alto: En la esquina superior izquierda se pueden observar cuatro alumnos con casi un 30 % de similitud entre sus trabajos, esto probablemente se deba a que tuvieron las mismas fuentes de información de algún libro o web (externo a este análisis). En el grupo inferior derecho hay similitudes bastante interesantes, porque superan el 50 %. Por ejemplo, el Alumno 31 presenta alrededor del 70 % de cercanía entre sus relaciones con los Alumnos 3, 7 y 25; esto se puede deber a que el Alumno 31 haya sido la fuente original del trabajo de sus otros tres compañeros porque los otros involucrados presentan relación entre ellos con un porcentaje más bajo que con el 31. Además el alto nivel de similitud probablemente se debió, según la explicación de los auxiliares del ramo, a que la primera tarea del IN72K era de investigación en Internet por lo cual mucha de la información puede provenir de algún sitio web al cual tuvieron acceso los alumnos (por ejemplo: Wikipedia).



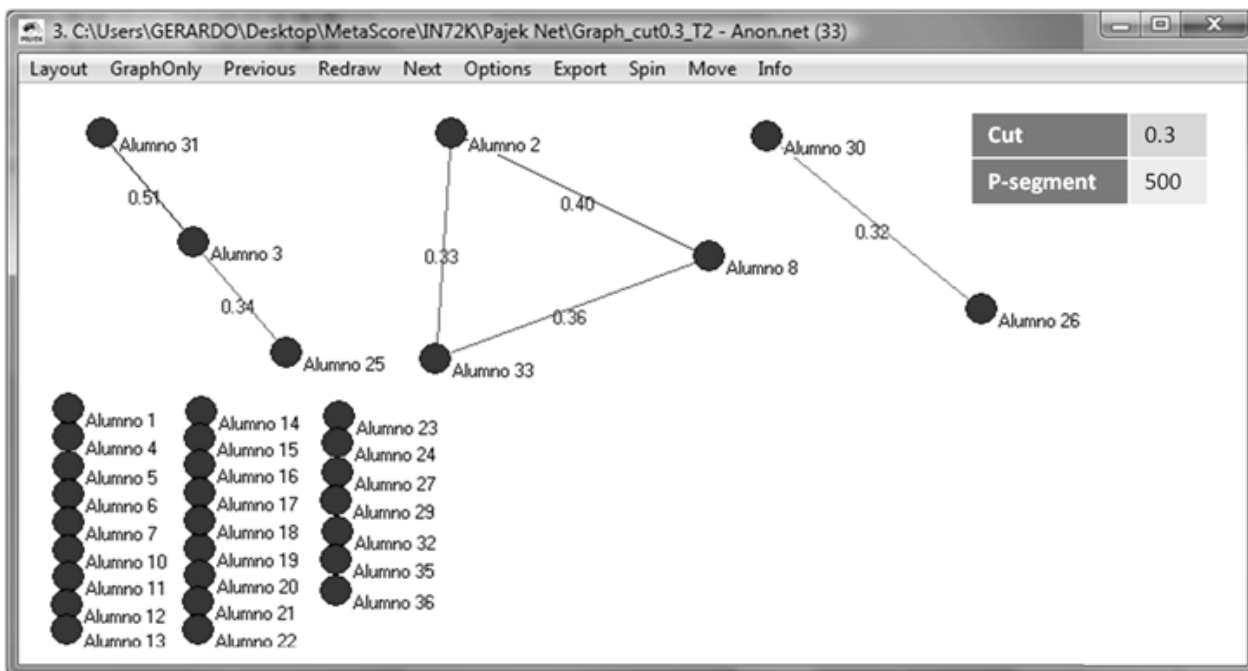


Figura 5.17: Análisis de Colusión para la tarea 2 del IN72K

En la tarea 2 (Figura 5.16) la similitud entre documentos disminuyó enormemente, respecto a la tarea anterior, esto se pudo deber a que ya no era un trabajo totalmente desarrollado para investigación web sino que presentaba partes analíticas. Nuevamente se observa que hay alta similitud entre el Alumno 31 y el Alumno 3 (como en la tarea 1), y esta nueva relación puede confirmar la hipótesis anterior “Que el Alumno 31 fuese la fuente de copia” ya que aquí se observa de forma bastante clara que el 31 sólo tiene cercanía con el Alumno 3 y nadie más, en cambio el Alumno 3 presenta cierta cercanía con el Alumno 25 por lo cual se puede suponer que el 3 copió de ambos compañeros. Además se reconocen otros dos conjuntos de similitud: el formado por los Alumnos 2, 8 y 33 con similitud entre todos. Y el formado por los Alumnos 26 y 30 que no tienen relación de similitud con algún otro grupo. En estos casos la similitud no es muy alta, por lo cual no se puede asegurar que exista grado de copia, es probable que los alumnos hayan copiado el enunciado de las tareas en sus documentos y por ello se agrega cierta tendencia de similitud. Los demás alumnos, no presentaron altos niveles de similitud, por ello se encuentran como nodos únicos y separados del grafo.

Cabe resaltar que se utiliza la definición “*análisis de similitud*” porque mediante el *Modelo de Fusión de Datos por Score* no se puede afirmar con absoluta certeza si existe copia entre documentos, o no, quedando este criterio al usuario quien, por contexto, tendrá un conocimiento experto respecto al entorno y la interacción entre Alumnos.

## Capítulo 6

# Conclusiones y trabajo futuro

Este capítulo está dividido en dos partes: La primera que aborda las conclusiones del trabajo de tesis, donde se mencionan las razones que hacen que el *Modelo de Fusión de Datos por Score* se considere como un método eficiente y confiable. La segunda parte es una breve reseña del trabajo futuro que se puede desencadenar como investigación posterior y que deriva del Modelo propuesto en esta tesis.

### 6.1. Conclusiones de la tesis

Con los resultados experimentales de los capítulos 4 y 5 se ha demostrado que el modelo propuesto en el Capítulo 3 es efectivo para determinar la probabilidad de copia entre Documentos y por ello ser un método eficiente de reconocimiento de copia escrita. Luego de finalizar el proyecto de tesis, se concluye:

- Se desarrolló un buen *Modelo de Fusión de Datos por Score* (MetaScore) que toma las mejores técnicas de linealización y combinación para entregar un resultado fiable y adecuado. Con esto se concluye que el objetivo principal de esta tesis fue logrado en su totalidad. Además, el desarrollo y experimentaciones con el Modelo Propuesto permitió alcanzar los demás objetivos de este proyecto de tesis (Capítulo 1, Sección 1.2. Página 4).
- Se realizó una investigación profunda respecto a técnicas de fusión de datos, haciendo un repaso en técnicas de ranking difuso que en una primera etapa de la tesis se consideró como la mejor manera de solucionar el problema, lo que luego se descartó al encontrar Métodos de Fusión de Score menos complejos y más eficientes.
- La inclusión del *Factor de credibilidad* en el *Modelo de Fusión de datos por Score propuesto* resultó otorgar una gran flexibilidad al Modelo, porque de este modo el usuario puede otorgar credibilidad a las *Técnicas de Detección de Similitud* de acuerdo a su criterio experto de las respuestas que estas

entregan. Es decir, podrá pre-validar si una *Técnica de Detección de Similitud* es confiable para un determinado conjunto de *Documentos Fuentes - Sospechosos*, o no.

- La propuesta del proceso DPD (Capítulo 4. Página 61) como metodología de evaluación del grado de copia otorga valor al *Modelo de Fusión*, porque indica cómo preparar los datos desde la *Base de datos inicial* hasta la lista de *Documentos Copiados*. Si se realizan futuras mejoras sobre el *Modelo propuesto*, se recomienda mantener el proceso DPD para analizar la copia entre documentos.
- Respecto a los indicadores de eficiencia, el *Modelo de Fusión de Datos por Score* presentó la mejor PRECISION (96.08 % en promedio). Superando por más de un 5 % a los modelos clásicos Z-Score–CombANZ, y con una desviación estándar con 5 % bajo el valor del clásico mencionado. Esto significa, de acuerdo a la formulación, que el *MetaScore* resalta los mejores scores. En otras palabras, el *Modelo Propuesto* permite encontrar eficientemente aquellos *Documentos con alta probabilidad de copia*.
- Así mismo, un indicador de RECALL no tan alto (78.09 % en promedio). Por debajo del 91.3 % del modelo clásico Estándar–CombANZ, quien a su vez tuvo un pésimo desempeño en su PRECISION con una media de 27.2 %, también es consecuente a la formulación del *Modelo de Fusión de Datos*. El *MetaScore* busca resaltar las duplas *Documento Fuente – Documento Sospechoso* y por ello se puede dar el caso que algunos resultados considerados *moderados* sean considerados como *no relevantes*, lo cual es válido.
- Para cerrar el tema de eficiencia. Contar con indicadores de PRECISION alto (96.08 %) y un RECALL medio-alto (78.09 %), es correcto para el objetivo del *Modelo de Fusión de Datos por Score*, porque al momento de hallar copia entre tareas, según la experiencia, en muchas ocasiones llevan el mismo enunciado'. Esto significa que al encontrarse tareas que contienen *enunciados* los *Métodos de Detección de Similitud* los reconocerán y les colocarán score dependiendo del tamaño del enunciado. Como este caso de *similitud de copia* sólo ocasionaría ruido en la detección final, al procesarse en el *Modelo de fusión de datos propuesto* el score asignado por el *método de similitud* bajará ,y luego de la evaluación con el punto de corte, la mencionada tarea con enunciado no será considerada como un probable *Documento fuente copia*.

## 6.2. Trabajo futuro

El trabajo futuro propuesto, se centra en la formulación del *Modelo de Fusión de Datos* particularmente en los tres primeros pasos: (a) Linealización, (b) Cálculo de la Frecuencia y (c) Cálculo del Valor de Información modificada.

### 6.2.1. Linealización

En el modelo se utilizó la linealización estándar en toda la formulación, como trabajo futuro se propone utilizar linealización por Sumatoria y por Z-score, para comprobar cual de estos métodos resulta más eficiente.

- **SUM.**

$$\overline{Sc}_k(x_i) = \frac{Sc_k(x_i) - ScMin_k}{ScSum_k - ScMin_k}; \quad x_i \in S_k; \forall k \in Mtd \quad (6.1)$$

Donde:

$$ScSum_k = \sum_{\forall x_i \in S_k} (Sc_k(x_i)) \quad (6.2)$$

Esta linealización tiene la característica que al utilizar la suma como valor máximo hará que ningún Score llegue a tener valor cercano a uno, ni mucho menos igual a uno. Puede que esto sea bueno, porque así la Función del Valor de la Información le otorgará un mejor puntaje a los pocos Scores que lleguen a tender a uno. Aumentando de ese modo la probabilidad que una copia escrita sea reconocida eficientemente. Como contraste, pueda que esta linealización haga que muchos documentos parcialmente copiados no sean reconocidos.

- **ZMUV.**

$$\overline{Sc}_k(x_i) = \frac{Sc_k(x_i) - \mu_k}{\sigma_k}; \quad x_i \in S_k; \forall k \in Mtd \quad (6.3)$$

Donde:

$$\mu_k = \frac{ScSum_k}{|Mtd|} \quad (6.4)$$

$$\sigma_k = \frac{\sum_{\forall x_i \in S_k} (Sc_k(x_i) - \mu_k)^2}{|Mtd|} \quad (6.5)$$

Con el Z-score, se espera reconocer con un alto valor de la información a aquellos scores que se alejan de la media  $\mu$ .

### 6.2.2. Cálculo de la Frecuencia

En el cálculo de la frecuencia se utilizó una división en p-segmentos, donde se contaba el número de elementos que caían dentro de un segmento en particular, luego se hallaba la frecuencia como la *Probabilidad* que el Score de una dupla *Documento Fuente – Documento Sospechoso* caiga en el segmento p-ésimo.

Se propone que el modelamiento de la probabilidad que el Score se encuentre en un rango determinado se calcule mediante una función de probabilidad y no mediante un análisis de frecuencia. Es decir, que se conozca el comportamiento de los scores para ciertos tipos de documentos y con ello se pueda modelar una función estocástica que indique la probabilidad de haya un score en el análisis.

$$P(Sc(x_i)) = f(x_i)$$

### 6.2.3. Cálculo del Valor de la Información modificada

Luego de la propuesta de la *Ecuación del Valor de la Información Modificada*, y sin un análisis matemático muy profundo se propone hacer una nueva versión de esta ecuación. Que esta vez incluya un “ranking” a partir de los Scores encontrados como se presenta en la Ecuación 6.6

$$IV_k(x_i) = -\log_2(F_k(r)) + (F_k(r))^{R_k(x_i)} \cdot \log_c |Mtd| \quad (6.6)$$

Donde:

$R_k(x_i)$  : Ranking dado por el método k-ésimo para el documento sospechoso  $x_i$

Además, para el valor del ranking se cumplirá:

$R_k(x_i) = -1$	Ecuación propuesta en el modelo
$R_k(x_i) = 0$	Ecuación original de Yu Suzuki et. al.[67]
$R_k(x_i) \geq 1$	Afina el Valor de la Información

Cuadro 6.1: Valores que puede tomar el ranking  $R_k(x_i)$

Que haciendo una ligera evaluación, considerando 6 p-segmentos y 100 elementos en análisis <sup>1</sup>.

<sup>1</sup>Por fines prácticos se en los resultados consideró que muchos de los elementos recibieron el mismo score.

Elementos	Score	p-segmento	Ranking	Yu Suzuki	Modelo tesis	Trabajo futuro
50	0.00001	$k_1$	30	0	0	0
30	0.005	$k_2$	10	0.0006528	0.0000779	0.0006490
10	0.01	$k_3$	6	0.0041141	0.0008725	0.0040900
6	0.1	$k_4$	3	0.0541986	0.0156345	0.0538941
3	0.9	$k_5$	2	0.6472532	0.2938813	0.6439398
1	1	$k_6$	1	1	1	1

Cuadro 6.2: Valor de la Información para los modelos de Yu Suzuki et. al. [67], propuesto en la tesis y el Trabajo Futuro

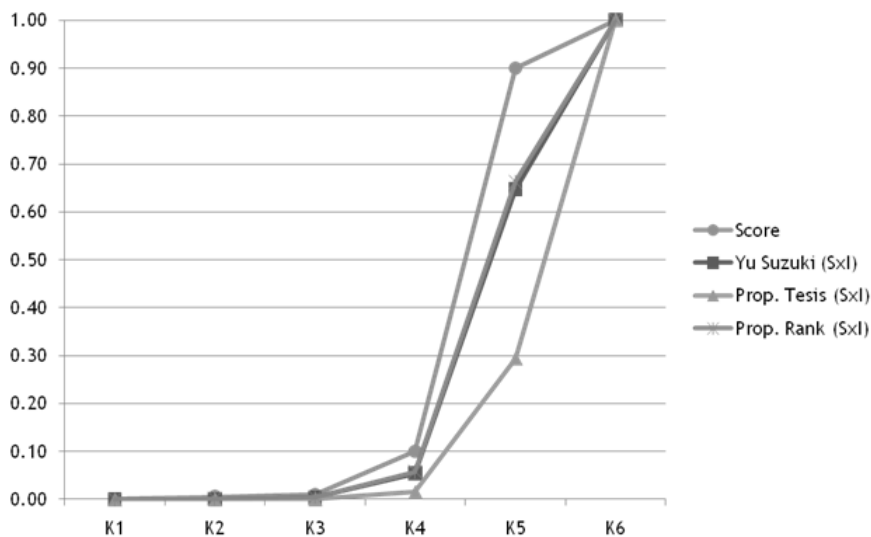


Figura 6.1: Curvas del Valor de la Información para los modelos de Yu Suzuki et. al [67], propuesto en la tesis y el Trabajo Futuro

# REFERENCIAS

- [1] Einat Amitay, David Carmel, Ronny Lempel, and Aya Soffer. Scaling ir-system evaluation using term relevance sets. In *SIGIR '04: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 10–17, New York, NY, USA, 2004. ACM.
- [2] Pauline Arnold and Paul Roberts. Plagiarism and collusion: Ignorance is no defence. *School of Psychology*, 1:112–123, 2001.
- [3] Armando Asti Vera. *Metodología de la investigación*. Kapelusz. Primera edición, 1973. Buenos Aires, Argentina.
- [4] Pedro Bádenas de la Peña. *Poesía Completa*. Alianza Editorial, 3era edición edition, 1983. Poesía de Kavafis, Konstantinos 1863-1933. Madrid, España.
- [5] John Baird. Current trends in college cheating. 17:515–522, 1980.
- [6] Alberto Barrón-cedeño, Paolo Rosso, David Pinto, and Juan Alfons. On cross-lingual plagiarism analysis using a statistical model. In *In: Proc. of PAN-08. (2008) in print*, 2008.
- [7] Peggy Bates and Margaret Fain. Detecting plagiarized papers, 2010. Coastal Carolina University.
- [8] M. M. Sufyan Beg and Nesar Ahmad. Soft computing techniques for rank aggregation on the world wide web. *World Wide Web*, 6(1):5–22, 2003.
- [9] Nicholas J. Belkin, Paul B. Kantor, Edward A. Fox, and Joseph A. Shaw. Combining the evidence of multiple query representations for information retrieval. *Inf. Process. Manage.*, 31:431–448, 1995.
- [10] Yaniv Bernstein and Justin Zobel. Accurate discovery of co-derivative documents via duplicate text detection. *Inf. Syst.*, 31:595–609, 2006.
- [11] Mikaela Björklund and Claes-Göran Wenestam. Academic cheating: frequency, methods and causes. In *Proceedings of the 1999 European Conference on Educational Research*, Lahti, Finland, 1999. European Conference on Educational Research.
- [12] Jude Carroll and Jon Appleton. Plagiarism: A good practice guide. Technical report, Oxford Brookes University, 2001.

- 
- [13] Web Intelligence Research Center and Universidad de Chile. DOCODE - DOcument COpy DEtector. Proyecto FONDEF. DII. Universidad de Chile.
- [14] Web Intelligence Research Center and Universidad de Chile. Proyecto FONDEF - D08I1015: Desarrollo e implementación de una herramienta computacional para la detección de copias en documentos digitales en la educación: DOcument COpy DEtector (DOCODE). Technical report, 2008.
- [15] Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, and Clifford Stein. *Introduction to Algorithms*. Number 16. The MIT Press, 2nd revised edition edition, September 2001.
- [16] Sara Cormeny. A word on plagiarism. *Washington Post*, 5:23–57, 1996.
- [17] John H. Coverdale and Marcus A. Henning. An analysis of cheating behaviors during training by medical students. *Medical Teacher*, 22(6):528–584, november 2000.
- [18] Marc Damashek. Gauging similarity with n-grams: Language-independent categorization of text. *Science*, 267(5199):843–848, February 1995.
- [19] Stephen Davis, Cathy Grover, Angela Becker, and Loretta McGregor. Academic dishonesty: Prevalence, determinants, techniques, and punishments. *Teaching of Psychology*, 19:16–20, 1992.
- [20] Stephen Davis and H. Wayne Ludvigson. Additional data on academic dishonesty and a proposal for remediation. *Teaching of Psychology*, 22:119–121, 1995.
- [21] Scott C. Deerwester, Susan T. Dumais, Thomas K. Landauer, George W. Furnas, and Richard A. Harshman. Indexing by latent semantic analysis. *Journal of the American Society of Information Science*, 41(6):391–407, 1990.
- [22] Louise Dennis. Student attitudes to plagiarism and collusion within computer science. volume 5, pages 217–221. University of Nottingham, 2004.
- [23] Persi Diaconis and R. L. Graham. Spearman’s footrule as a measure of disarray. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(2):262–268, 1977.
- [24] Didier Dubois and Henri Prade. *Fuzzy sets and systems - Theory and applications*, volume 2. Academic press, New York, 1980.
- [25] Cynthia Dwork, Ravi Kumar, Moni Naor, and D. Sivakumar. Rank aggregation methods for the web. In *WWW '01: Proceedings of the 10th international conference on World Wide Web*, pages 613–622, New York, NY, USA, 2001. ACM.
- [26] Sven M. Eissen and Benno Stein. Intrinsic plagiarism detection. In *Proceedings of the European Conference on Information Retrieval (ECIR-06)*, pages 565–569. Springer, 2006.
- [27] Real Academia Española. Diccionario de la lengua española - vigésima segunda edición. versión on-line.
- [28] Bordignon. Fabian., Ricardo. Tolosa, Carlos. Rodríguez, and José Peri. Primeras experiencias en la detección de plagio en el ambiente educativo. *Primera Jornada de Educación en Informática y TICS en Argentina*, pages 97–104, 2003.



- 
- [29] Barbara J. Fly, William P. van Bark, Laura Weinman, Karen Strohm, and Patrick R. Lang. Ethical transgressions of psychology graduate students: Critical incidents with implications for training. *Professional Psychology: Research and Practice*, 28(5):492–495, 1997.
- [30] Arlene Franklyn-Stokes and Stephen E. Newstead. Undergraduate cheating: Who does what and why? *Studies in Higher Education*, 20(2):159–172, 1995.
- [31] Dan Gusfield. *Algorithms on Strings, Trees, and Sequences - Computer Science and Computational Biology*, volume 4. Cambridge University Press, 1997.
- [32] Mavis Hetherington and Solomon Feldman. College cheating as a function of subject and situational variables. *Journal of Educational Psychology*, 55:212–218, 1964.
- [33] Michael Josephson and Melissa Mertz. A resource to help teachers and administrators promote integrity and prevent academic dishonesty. 1:52–56, 2004.
- [34] Alex Kellogg. Students plagiarize less than many think: A new study finds. *The Chronicle of Higher Education*, 2:84–102, 2002.
- [35] Maurice G. Kendall. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93, 1938.
- [36] Philip C. Kolin. *Successful writing at work*. Concise ed. - Boston : Houghton Mifflin, Boston, MA, USA, 2006.
- [37] Thomas K. Landauer and Susan T. Dumais. A solution to plato’s problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104(2):211–240, April 1997.
- [38] Joon H. Lee. Analyses of multiple evidence combination. In *SIGIR '97: Proceedings of the 20th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 267–276, New York, NY, USA, 1997. ACM.
- [39] Joon Ho Lee. Combining multiple evidence from different relevant feedback networks. In *Proceedings of the Fifth International Conference on Database Systems for Advanced Applications (DASFAA)*, pages 421–430. World Scientific Press, 1997.
- [40] Michael D. Lee, Brandon Pincombe, and Matthew Welsh. A comparison of machine measures of text document similarity with human judgments. In *CogSci2005*, pages 1254–1259. Erlbaum, 2005.
- [41] Alberta Lipson and Norma McGraven. Undergraduate academic dishonesty at mit. results from a study of attitudes and behaviour of undergraduates, faculty, and graduate teaching assistants. *MIT*, 1993.
- [42] Yu-Ting Liu, Tie-Yan Liu, Tao Qin, Zhi-Ming Ma, and Hang Li. Supervised rank aggregation. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*, pages 481–490, New York, NY, USA, 2007. ACM.
- [43] Vincenzo Loia, Masoud Nikravesh, and Lotfi A. Zadeh. Fuzzy logic and the internet. *Soft Comput.*, 6(5):285–286, 2002.

- 
- [44] George MacDonald Ross. Plagiarism in philosophy: Prevention better than cure. *Discourse*, 3(2):23–57, 2003.
- [45] John Makhoul, Francis Kubala, Richard Schwartz, and Ralph Weischedel. Performance measures for information extraction. In *In Proceedings of DARPA Broadcast News Workshop*, pages 249–252, 1999.
- [46] Brian Martin. Plagiarism: a misplaced emphasis. *Journal of Information Ethics*, 3(2):36–47, 1994.
- [47] Donald L. McCabe, Linda K. Treviño, and Kenneth D. Butterfield. Cheating in academic institutions: A decade of research. *ETHICS AND BEHAVIOR*, 11(3):219–232, 2001.
- [48] Mark Montague. *MetaSearch: Data Fusion for Document Retrieval*. PhD thesis, Dartmouth College, March 2002.
- [49] Mark Montague and Javed A. Aslam. Relevance score normalization for metasearch. In *CIKM '01: Proceedings of the tenth international conference on Information and knowledge management*, pages 427–433, New York, NY, USA, 1997. ACM.
- [50] Jenny Moon. Academic honesty, plagiarism and cheating: a self-instruction unit for level 3 students, 2005.
- [51] Kwong Bor Ng and Paul B. Kantor. An investigation of the preconditions for effective data fusion in information retrieval: A pilot study. In *CProceedings of the 61st Annual Meeting of the American Society for Information Science*, pages 166–178, 1998.
- [52] Mario Núñez M. Plagio en la era digital. *Discourse UPRM*, 1:65–88, 2006.
- [53] Gabriel Oberreuter, Gaston L’Huillier, Sebastián A. Ríos, and Juan D. Velásquez. Fastdocode: Finding approximated segments of n-grams for document copy detection - lab report for pan at clef 2010. In *CLEF (Notebook Papers/LABs/Workshops)*, 2010.
- [54] American Association of University Professors. American association of university professors, 2008. versión on-line.
- [55] Authorship PAN: Uncovering Plagiarism and Social Software Misuse. Overview of the 1st international competition on plagiarism detection, 2010.
- [56] Timothy J. Ross. A review of: Fuzzy logic with engineering applications. page 600, New York, NY. United States, 1995. McGraw-Hill.
- [57] Comas Rubén, Sureda Jaume, Ortega D., and Santos Urbina. Ciber-plagio académico: la generación de cortar y pegar, 2006. III Congreso On line: Ciber sociedad.
- [58] Gerard Salton. *Automatic text processing: the transformation, analysis, and retrieval of information by computer*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1989.
- [59] SciELO. Plagio académico, ¿robo y fraude? *Información tecnológica*, 19:1–10, 2008.
- [60] Leanne Seaward and Stan Martin. Intrinsic plagiarism detection using complexity analysis. In *In: Proc. of PAN-09. (2009) in print*, 2009.

- 
- [61] Erik Selberg and Oren Etzioni. On the instability of web search engines. In *In Proceedings of RIAO '00*, pages 223–235, 2000.
- [62] Claude E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27:379–423, 623–656, July, October 1948.
- [63] Joseph A. Shaw and Edward A. Fox. Combination of multiple searches. In *Text REtrieval Conference*, pages 243–252, 1993.
- [64] Masamichi Shimura. Fuzzy sets concept in rank-ordering objects. *Journal of Mathematical Analysis and Applications*, 43(3):717–733, 1973.
- [65] Antonio Si, Hong Va Leong, and Rynson W. H. Lau. Check: a document plagiarism detection system. In *SAC '97: Proceedings of the 1997 ACM symposium on Applied computing*, pages 70–77, New York, NY, USA, 1997. ACM.
- [66] Ian Soboroff, Charles Nicholas, and Patrick Cahan. Ranking retrieval systems without relevance judgments. In *SIGIR '01: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 66–73, New York, NY, USA, 2001. ACM.
- [67] Yu Suzuki, Kenji Hatano, Masatoshi Yoshikawa, Shunsuke Uemura, and Kyoji Kawagoe. A relevant score normalization method using shannon's information measure. In *ICADL*, pages 311–316, 2005.
- [68] Northwestern University The Writing Place. Avoiding plagiarism, 2005.
- [69] Donald Thistlethwaite. The impact of the episodes of may, 1970 upon american university students. *Research in Higher Education*, 1:225–243, 1973. 10.1007/BF00991531.
- [70] Marlin Thomas and Samuel Rudin. Plagiarism detection software. Technical report, Iona College, 2008.
- [71] Jean Tirole. Colusión y la teoría de organizaciones. *Procedimientos del sexto congreso del mundo de la sociedad econométrica*, 2:151–206, 1992.
- [72] Merijn Van Erp and Lambert Schomaker. Variants of the borda count method for combining ranked classifier hypotheses. In *Seventh International Workshop on Frontiers in Handwriting Recognition*, pages 443–452, 2000.
- [73] Cornelis Joost van Rijsbergen. Information retrieval. *Annual Review of Information Science and Technology*, 2, 1979.
- [74] Christopher C. Vogt and Garrison W. Cottrell. Fusion via a linear combination of scores. *Information Retrieval*, 1:151–173, 1999. 10.1023/A:1009980820262.
- [75] John Walker. Student plagiarism in universities: what are we going to do about it? pages 89–106, 1998.
- [76] Shengli Wu and Fabio Crestani. Methods for ranking information retrieval systems without relevance judgments. In *SAC '03: Proceedings of the 2003 ACM symposium on Applied computing*, pages 811–816, New York, NY, USA, 2003. ACM.

- 
- [77] Shengli Wu, Fabio Crestani, and Yaxin Bi. Evaluating score normalization methods in data fusion. In *AIRS*, pages 642–648, 2006.
- [78] Shengli Wu and Sally McClean. Performance prediction of data fusion for information retrieval. *Inf. Process. Manage.*, 42(4):899–915, 2006.
- [79] Shengli Wu, Qili Zhou, Yaxin Bi, and Xiaoqin Zeng. Performance weights for the linear combination data fusion method in information retrieval. In Aijun An, Stan Matwin, Zbigniew Ras, and Dominik Slezak, editors, *Foundations of Intelligent Systems*, volume 4994 of *Lecture Notes in Computer Science*, pages 465–470. Springer Berlin / Heidelberg, 2008.
- [80] Ronald R. Yager. On ordered weighted averaging aggregation operators in multicriteria decisionmaking. *IEEE Trans. Syst. Man Cybern.*, 18(1):183–190, 1988.
- [81] H. Peyton Young. An axiomatization of borda’s rule. *Journal of Economic Theory*, 9:43–52, 1974.

# Apéndices

## Apéndice A

# Conceptos de teoría de la información

Este apéndice aborda temas de la *Teoría de la Información*, se hace un repaso del concepto de Información, detallando el modelo de comunicación y posteriormente los conceptos estadísticos de entropía.

### A.1. Información

El estudio de la *Teoría de la Información* fue iniciado por el investigador Claude E. Shannon en su trabajo *A Mathematical Theory of Communication* del año 1948 [62] y parte desde el concepto de la *Comunicación*, que en el contexto de la *Información* es empleado en un sentido muy amplio en el que “quedan incluidos todos los procedimientos mediante los cuales una mente puede influir en otra”. Es decir, se pueden considerar todas las formas que el ser humano usa para transmitir ideas: la palabra hablada, escrita, los gestos, la música, las imágenes, los movimientos, etc. y todos sus medios de transmisión: teléfono, radio, internet, etc. El modelo propuesto por Shannon es un sistema general de la comunicación que parte de una fuente de información, desde la cual, a través de un transmisor se emite una señal, la cual viaja por un canal, pero a lo largo de su viaje puede ser interferida por algún ruido. La señal sale del canal, la cual llega a un receptor que decodifica la información, convirtiéndola posteriormente en mensaje que pasa a un destinatario (Ver Figura A.1).

Cabe resaltar que en este contexto no es importante el *significado del mensaje* sino que el interés principal está en lo relacionado a la capacidad y fidelidad para enviar información en los diferentes medios de transmisión. Es decir, no se centra en la *cantidad* sino en la *calidad* de la información.

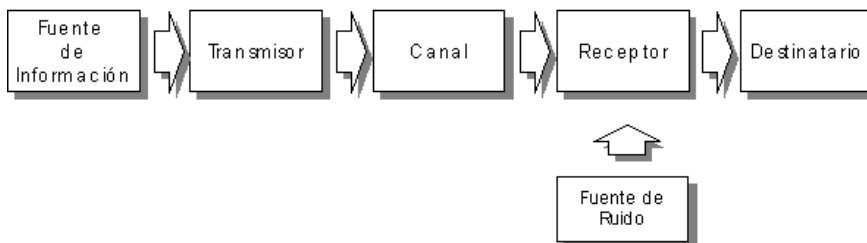


Figura A.1: Modelo del sistema de comunicación

## A.2. Teoría de la Información

En la *Teoría de la Información* el concepto de *Información* es definido en términos estrictamente estadísticos, y se relaciona con la libertad de elección que tenemos para seleccionar un mensaje determinado de un conjunto de posibles mensajes. Este concepto supone la existencia de probabilidad o incertidumbre. La incertidumbre implica que existen diferentes alternativas a elegir y que queda a criterio del receptor encontrar aquella elección que otorgue mayor *Valor de la Información*.

Se debe asumir que ambos extremos del canal de comunicación: la fuente y el receptor manejan el mismo código. Al cuantificar la *información* proporcionada por la fuente al receptor, se puede asociar una probabilidad o incertidumbre, siendo algunos mensajes más probables que otros.

Por ejemplo, sea una caja con 5 *pelotitas negras* y 1 *pelotita blanca*, todas con la misma probabilidad de ser elegidas, o sea  $1/6$ . Si consideramos que la *información* significa conocer el orden en que se puedan sacar las *pelotitas*, la incertidumbre dependerá de la *pelotita blanca* porque una vez que salga esta *pelotita* de la caja se podrá estar totalmente seguro, con probabilidad igual a 1, que las demás *pelotitas* que salgan serán necesariamente negras. Hagamos un supuesto de extracción de *pelotitas* de la caja: Si en la primera extracción sale una *pelotita negra* en la segunda extracción puede salir otra negra o la blanca (incertidumbre). Si en la segunda extracción sale la *pelotita blanca* entonces se puede asegurar que en las siguientes 4 extracciones restantes sólo saldrán *pelotitas negras* (certeza). Entonces se puede decir que el evento “extracción de la bolita blanca” es el que entregue mayor *Valor de la Información* porque será el desencadenante para conocer a priori los resultados futuros. A este concepto, se le denomina *Información* y Shannon lo define matemáticamente como:

$$I(x_i) = -\log_a P(x_i) \quad (\text{A.1})$$

Donde:

- $I(x_i)$  : Información de una variable aleatoria.
- $x_i$  : Variable aleatoria discreta.
- $P(x_i)$  : Probabilidad de ocurrencia de  $x_i$ .

El logaritmo en base  $a$  usualmente es un  $\log_2$  porque en entornos computacionales se suelen utilizar bits, es decir 1s y 0s.

### A.3. Entropía

La *Entropía* puede ser considerada como una medida de incertidumbre, y su importancia radica en que permite reducir la incertidumbre. En otras palabras, indica en qué medida es *no predecible* un resultado experimental cuando está sujeto a incertidumbre. De acuerdo a Wikipedia<sup>1</sup>:

“La entropía también puede ser entendida como la cantidad de información promedio que contienen los símbolos usados. Los símbolos con menor probabilidad son los que aportan mayor información; por ejemplo, si se considera como sistema de símbolos a las palabras en un texto, palabras frecuentes como “que”, “el”, “a” aportan poca información, mientras que palabras menos frecuentes como “corren”, “niño”, “perro” aportan más información (si de un texto dado borramos un “que”, seguramente no afectará a la comprensión y se sobreentenderá, no siendo así si borramos la palabra “niño” del mismo texto original). Cuando todos los símbolos son igualmente probables, todos aportan información relevante y la entropía es máxima.

Con la definición de Shannon, la entropía se puede definir como la cantidad de ruido o desorden que agrega el canal en el Modelo del sistema de comunicación (Ver Figura A.1). Y se expresa:

$$H(X) = - \sum_{i=1}^n P(x_i) \log_a P(x_i) \quad ; \quad i \in [1; n] \quad (\text{A.2})$$

---

<sup>1</sup><http://es.wikipedia.org>



## Apéndice B

# Indicadores de Eficiencia en Recuperación de la información

Existen diferentes indicadores para evaluar la eficiencia de un sistema, este apéndice presenta los indicadores de eficiencia utilizados en el proyecto de tesis: Precision, Recall, Accuracy y F-measure. Estos indicadores manejan 2 conceptos: La *relevancia* y el *reconocimiento*. La *relevancia* se refiere a si, en un análisis supervisado, el resultado fue reconocido como importante; en el caso del plagio, relevancia es si un Documento es efectivamente copiado. Por otro lado, el *reconocimiento* está orientado al resultado del sistema a evaluar y si dicho resultado indicó que era importante; en el caso del plagio, reconocimiento es si el sistema reconoció que el Documento era una copia. Por estos conceptos se crea una matriz de confusión B.1.

	<b>Relevante</b>	<b>No-relevante</b>
<b>Reconocido</b>	Verdadero positivo (VP)	Verdadero negativo (VN)
<b>No reconocido</b>	Falso negativo (FN)	Falso positivo (FP)

Cuadro B.1: Matriz de confusión

De la matriz de confusión se extraen nuevos conceptos: **(a) Verdadero positivo.** Cuando el Sistema reconoció al resultado como relevante y verdaderamente el resultado era relevante (acertó). **(b) Verdadero negativo.** Cuando el Sistema reconoció al resultado como relevante pero en la realidad el resultado no era relevante (falló). **(c) Falso negativo.** Cuando el sistema no reconoció al resultado como relevante, sin embargo el resultado si era relevante (falló). **(d) Falso positivo.** Cuando el sistema no reconoció al resultado como relevante y efectivamente el resultado no era relevante (acertó). Con estos conceptos se desarrollan los indicadores.

## B.1. Precision

El indicador de *Precision* expresará “De todos los resultados que sistemas dijo que eran relevantes, cuantos efectivamente eran relevantes”.

$$Precision = \frac{VP}{VP + FP} \quad (B.1)$$

Donde:

- Precision* → 0 : El sistema indicó pocos que si eran relevantes  
*Precision* → 1 : El sistema indicó muchos que si eran relevantes

## B.2. Recall

El *Recall* indica “De todos los resultados relevantes que existían, a cuántos de ellos reconoció el sistema como relevantes”.

$$Recall = \frac{VP}{VP + FN} \quad (B.2)$$

Donde:

- Recall* → 0 : El sistema no reconoció muchos como relevantes  
*Recall* → 1 : El sistema reconoció muchos relevantes

## B.3. Accuracy

El indicador del *Accuracy* muestra la certeza que se tuvo para reconocer los resultados relevantes.

$$Accuracy = \frac{VP + VN}{VP + VN + FP + FN} \quad (B.3)$$

En la siguiente figura, se muestra la diferencia entre *Accuracy* y *Precision*. A la izquierda se observa un caso de alto *Accuracy* y baja *Precision*, hace que *en promedio* todos los puntos se encuentren en el centro.

Mientras que a la derecha está el caso de bajo Accuracy y alto Precision, hace que todos los puntos tiendan en un mismo punto, aunque debido a la falta de Accuracy el puntaje no es acertado en el centro.

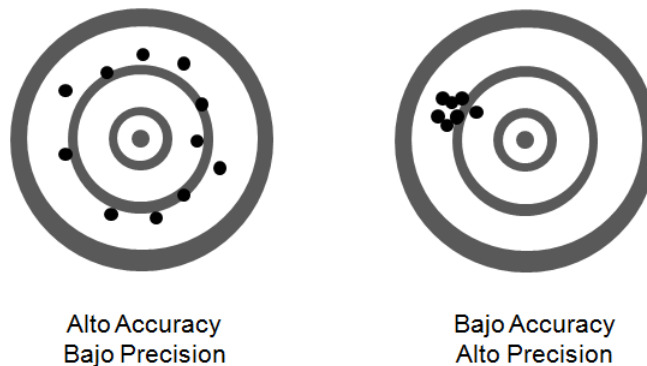


Figura B.1: Diferencia entre Accuracy y Precision

## B.4. F-measure

Finalmente un indicador que incluye en su función al Precision y al recall. Un gran valor de Precision hace el Recall disminuya y un gran Recall hace que el Precisión decrezca. El F-measure es la media armónica entre Precision y Recall:

$$Accuracy = \frac{(1 + \beta^2) \cdot Precision \cdot Recall}{\beta^2 \cdot Precision + Recall} \quad (B.4)$$

Donde:

$\beta$	Comentario
= 1	Igual peso para Precision y Recall
> 1	Mayor peso para el Recall
< 1	Mayor peso para el Precision

## Apéndice C

# Análisis de la técnica de linealización de scores de Yu Suzuki et. al.

Para la técnica de linealización de scoring de Yu Suzuki et. al. [67] se deben cumplir tres pasos (Las notaciones y expresiones fueron presentadas y explicadas en la sección 2.4.3):

### 1. Linealización de scores.

$$\overline{Sc}_k(x_i) = \frac{Sc_k(x_i) - ScMin_k}{ScMax_k - ScMin_k}; \quad x_i \in S_k; \forall k \in Mtd \quad (C.1)$$

### 2. Cálculo del valor de la información.

Se necesita pre-calcular el número de elementos  $E(Seg_k(r))$  que pertenecen a un segmento en particular  $Seg_k(r)$ :

$$E(Seg_k(r)) = |\{x_i | \overline{Sc}_k(x_i) \in Seg_r; \forall x_i \in Src\}|; \quad \forall k \in Mtd \quad (C.2)$$

Donde:

$$Seg_k(r) = \left[ \frac{r-1}{p}; \frac{r}{p} \right]; r = 1, 2, \dots, p-1, p; \forall k \in Mtd \quad (C.3)$$

Además el valor de  $p$  es ingresado por criterio personal del usuario, siendo definido como

$p$  : Número de divisiones que se requiere repartir el *rango*  $\in [0; 1]$ ;  $p \in \mathbb{Z}^+$

Posteriormente calcular la probabilidad que hayan  $E(Seg_k(r)) = d$  documentos en un segmento  $r$  como la frecuencia  $F_k(r)$ :

$$P \left( E(Seg_k(r))=d / \overline{Sc}_k(x_i) \in Seg_k(r) \right) = F_k(r) = \frac{E(Seg_k(r))}{|Src|}; \quad \forall k \in Mtd \quad (C.4)$$

Calcular el *valor de la información*.

$$IV_k(x_i) = -\log_2(F_k(r)); \quad \forall x_i \in Src; \forall k \in Mtd \quad (C.5)$$

Y finalmente linealizar el valor de la información.

$$\overline{IV}_k(x_i) = \frac{IV_k(x_i) - IVMin_k}{IVMax_k - IVMin_k}; \quad x_i \in S_k; \forall k \in Mtd \quad (C.6)$$

### 3. Integración de scores.

$$\widetilde{Sc}_k(x_i) = \overline{Sc}_k(x_i) \cdot \overline{IV}_k(x_i); \quad \forall x_i \in Src; \forall k \in Mtd \quad (C.7)$$

Aplicando el método de linealización de Suzuki et. al. para un grupo de valores de prueba (*Test 1*). Sea  $k$  un sistema de detección de similitud de documentos (Donde:  $k \in Mtd$ ) y  $p$  el número de segmentos que se desea dividir el rango linealizado después del *Paso 1* (Donde:  $rango \in [0; 1]$ ). Para el test será  $p = 6$ :

SEGMENTOS $Seg_k(r)$	ELEMENTOS $E(Seg_k(r))$	SCORE ORIGINAL $\overline{Sc}_k(x_i)$	PROBABILIDAD $F_k(r)$	VALOR INFORMACIÓN $IV_k(x_i)$	LIN. VALOR INF $\overline{IV}_k(x_i)$	LIN. SUZUKI $\widetilde{Sc}_k(x_i)$
$r = 1$	50	0,00001	0,5	1	0	0
$r = 2$	30	0,005	0,3	1,736965594	0,130578379	0,000652892
$r = 3$	10	0,01	0,1	3,321928095	0,41140809	0,004114081
$r = 4$	6	0,1	0,06	4,058893689	0,541986469	0,054198647
$r = 5$	3	0,9	0,03	5,058893689	0,719170289	0,64725326
$r = 6$	1	1	0,01	6,64385619	1	1

Cuadro C.1: Test 1 de la técnica de linealización Yu Suzuki et. al.

Como se observa, se tiene un total de 100 documentos analizados para este  $k$ -ésimo método de detección de similitud:

$$TOTAL_k = \sum_{r=1}^p E(Seg_k(r)) = 100$$

Además se considera que el score es único para cada segmento. Es decir, del Cuadro C.1: Los 50 elementos del segmento  $r = 1$  tienen un  $Score = 0,00001$ ; Mientras que los 10 elementos del segmento  $r = 3$  poseen un  $Score = 0,01$ .

La representación gráfica de estos resultados se aprecia en la Figura C.1, donde se puede notar que el método de Suzuki et. entrega una mejor linealización. Por ejemplo (de la Figura C.1): En el quinto segmento ( $r = 5$ ), el Score por linealización estándar otorga a los elementos un valor de 0,9, en cambio la técnica de Suzuki le calcula un valor de 0,647 porque el valor de la información del  $r = 5$  no es el más alto de los segmentos (En esta prueba, la sexta división presenta un mayor *Valor de la información* que cualquier otro. Ver Cuadro C.1).

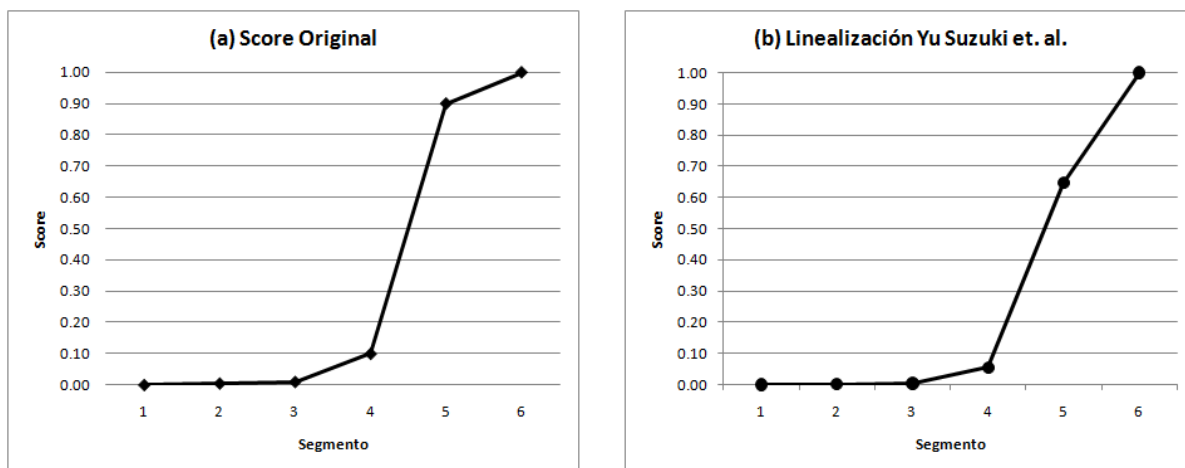


Figura C.1: Linealización para el método k-ésimo: (a) Score original (b) Score Suzuki et. al.

Haciendo una segunda prueba (*Test 2*) para probar la efectividad de la linealización de Suzuki, se tienen los valores del Cuadro C.2. En estos valores se aprecia que la cantidad de elementos que caen dentro de un determinado segmento  $r$  (Donde:  $r = 1, 2, \dots, p$ ) es bastante peculiar:

$$E(\text{Seg}_k(r)) = 3, \forall r \neq 5$$

$$E(\text{Seg}_k(5)) = 30$$

Esto hace que el *valor de la información* juegue un rol importante dentro del cálculo de la linealización de scores. En el Cuadro C.2, la quinta columna indica el *valor de la información* obtenido luego de aplicar la Ecuación C.5 y se observa que el valor que respecta al quinto segmento ( $r = 5$ ) es más bajo que cualquier otro segmento ( $r = 1, 2, 3, 4, 6$ ); luego en la sexta columna se observa que en la linealización del Valor de la información el  $r = 5$  es cero.<sup>1</sup>

SEGMENTOS $\text{Seg}_k(r)$	ELEMENTOS $E(\text{Seg}_k(r))$	SCORE ORIGINAL $\overline{Sc}_k(x_i)$	PROBABILIDAD $F_k(r)$	VALOR INFORMACIÓN $IV_k(x_i)$	LIN. VALOR INF $\overline{IV}_k(x_i)$	LIN. SUZUKI $\widetilde{Sc}_k(x_i)$
$r = 1$	3	0,00001	0,066666667	3,906890596	1	0,00001
$r = 2$	3	0,005	0,066666667	3,906890596	1	0,005
$r = 3$	3	0,01	0,066666667	3,906890596	1	0,01
$r = 4$	3	0,1	0,066666667	3,906890596	1	0,1
$r = 5$	30	0,9	0,666666667	0,584962501	0	0
$r = 6$	3	1	0,066666667	3,906890596	1	1

Cuadro C.2: Test 2 de la técnica de linealización Yu Suzuki et. al.

Para el caso del Cuadro C.2, al igual que el caso del Cuadro C.2 se tiene un  $p = 6$  segmentos; Además suma un total de 45 documentos analizados.

<sup>1</sup>En este caso particular, la linealización del valor de la información ( $\overline{IV}_k(x_i)$ ) se observa binaria, esto se debe a la forma de los datos de ingreso, porque como sólo existen dos cantidades de elementos  $E(\text{Seg}_k(r))$  para los segmentos también existirán sólo dos valores de  $\overline{IV}_k(x_i)$

$$TOTAL_k = \sum_{r=1}^p E(Seg_k(r)) = 45$$

Este nuevo test entrega las siguientes gráficas:

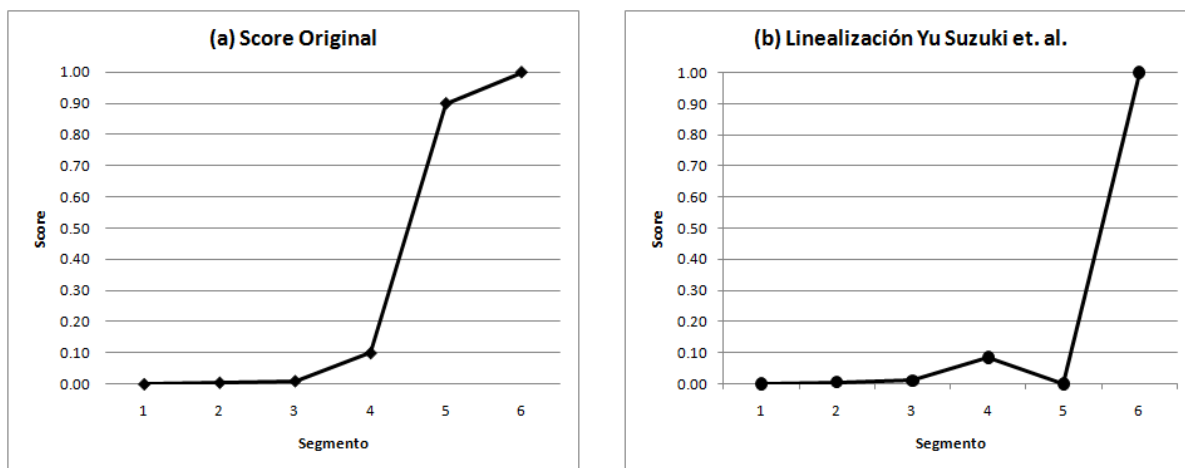


Figura C.2: Linealización para el método k-ésimo: (a) Score original (b) Score Suzuki et. al.

Nótese como cambia el score en la Gráfica C.2 (b), en los casos donde el número de elementos  $E(Seg_k(r))$  es igual para los segmentos  $r$ , luego del proceso de linealización de Suzuki ( $\widetilde{S}_{C_k}(x_i)$ ), estos mantienen su *Score original*  $\widetilde{S}_{C_k}(x_i)$ . Mientras que para el caso de  $r = 5$  el valor  $\widetilde{S}_{C_k}(x_i)$  es igual a cero 0. Esto mismo se puede observar en la columna 7 del Cuadro C.2. Es debido a este comportamiento que la técnica de Suzuki alcanza buenos resultados, porque considera “la importancia” de los documentos de acuerdo su valor de la información.

Como una nota particular. El valor de la información disminuirá, si existen muchos documentos en un segmento  $r$ ; y viceversa si existen pocos documentos en  $r$ . Matemáticamente se puede expresar como a continuación.

Si:

$$\frac{E(Seg_k(r))=n}{n \rightarrow \infty} \Rightarrow VI \rightarrow 0$$

$$\frac{E(Seg_k(r))=n}{n \rightarrow 0} \Rightarrow VI \rightarrow \infty$$

Finalmente, como se esperaba, la linealización por la *técnica de Yu Suzuki* entregó mejores resultados respecto a la *linealización estándar*. Siendo comprobado que al agregar el *valor de la información* a la linealización los resultados se vuelven más fiables y por ende otorga un valor agregado que las técnicas presentadas por Montague y Aslam en [49].

## Apéndice D

### Tablas de valores resultantes

En este apéndice se incluyen todos los resultados en formatos obtenidos durante la experimentación (Sección 4) y análisis de resultados (Sección 5), que por su dimensión no fueron agregados directamente al capítulo respectivo.

#### D.1. Lista de Documentos seleccionados del PAN Corpus 2010

Lista de Documentos	
suspicious-document04112	suspicious-document01145
suspicious-document13327	suspicious-document01806
suspicious-document05387	suspicious-document02542
suspicious-document02415	suspicious-document03238
suspicious-document09248	suspicious-document03980
suspicious-document01491	suspicious-document04724
suspicious-document06601	suspicious-document05453
suspicious-document11262	suspicious-document06100
suspicious-document01203	suspicious-document06706
suspicious-document05259	suspicious-document07449
suspicious-document09317	suspicious-document08227
suspicious-document12515	suspicious-document08870
suspicious-document01001	suspicious-document09646
suspicious-document04061	suspicious-document10350
suspicious-document07564	suspicious-document11082

Del cuadro D.2, se tiene un total de 266 *Documentos Fuentes* para los 80 *Documentos Sospechosos*. Luego se eligieron otro 236 *Documentos Fuentes* que no pertenecen a ningún *Documento Sospechoso*. Con ello se logró contar con: 80 *Documentos Sospechosos* y 500 *Documentos Fuentes*.



Lista de Documentos	
suspicious-document10297	suspicious-document11741
suspicious-document13596	suspicious-document12534
suspicious-document01772	suspicious-document13370
suspicious-document03536	suspicious-document14194
suspicious-document05160	suspicious-document00373
suspicious-document07258	suspicious-document00958
suspicious-document08828	suspicious-document01542
suspicious-document10320	suspicious-document02151
suspicious-document11966	suspicious-document02841
suspicious-document13588	suspicious-document03525
suspicious-document00256	suspicious-document04104
suspicious-document01258	suspicious-document04697
suspicious-document02265	suspicious-document05355
suspicious-document03320	suspicious-document05993
suspicious-document04306	suspicious-document06638
suspicious-document05385	suspicious-document07363
suspicious-document06694	suspicious-document08124
suspicious-document07854	suspicious-document08734
suspicious-document09024	suspicious-document09428
suspicious-document10044	suspicious-document10151
suspicious-document11137	suspicious-document10784
suspicious-document12012	suspicious-document11568
suspicious-document12876	suspicious-document12292
suspicious-document13920	suspicious-document13085
suspicious-document00434	suspicious-document14418

Cuadro D.1: Lista de los 80 Documentos seleccionados desde el PAN Corpus 2010

## D.2. Análisis para elección del punto de corte óptimo

Para encontrar el punto de corte óptimo, se hicieron varias iteraciones hasta llegar al valor cercano que entregaba el máximo F-measure.

### D.2.1. Corte entre 0 y 1

Para este corte se tomaron los siguientes valores para el algoritmo greedy:

Con lo que se obtiene:

**D.2.2. Corte entre 0 y 0.004**

Para este corte se tomaron los siguientes valores para el algoritmo greedy:

Con lo que se obtiene:

**D.2.3. Corte entre 0.00022 y 0.00042**

Para este corte se tomaron los siguientes valores para el algoritmo greedy:

Con lo que se obtiene:

**D.2.4. Corte entre 0.0003122 y 0.0003922**

Para este corte se tomaron los siguientes valores para el algoritmo greedy:

Con lo que se obtiene:

### **D.3. Análisis para elección del p-segmento óptimo**

#### **D.3.1. p-segmento entre 2 y 1000**

Para este corte se tomaron los siguientes valores para el algoritmo greedy:

Con lo que se obtiene:

#### **D.3.2. p-segmento entre 400 y 600**

Para este corte se tomaron los siguientes valores para el algoritmo greedy:

Con lo que se obtiene:

#### **D.4. Indicadores para todos los documentos seleccionados**

Luego de aplicar el *Modelo de Fusión de Scores* y conociendo los valores de corte y p-segmentos óptimos de  $3,5 \cdot 10^{-4}$  y 500 respectivamente. Se obtuvieron los siguientes indicadores para los 80 Documentos de la base seleccionada.

Además, se calculo la *media* y la *desviación estándar* de los documentos analizados (ver al final del cuadro)

<b>Fuentes</b>	<b>Nombre Sospechoso</b>	<b>Nombre Fuentes</b>	
28	suspicious-document04112.txt	source-document06986.txt source-document04818.txt source-document05904.txt source-document07121.txt source-document05928.txt source-document00527.txt source-document00257.txt source-document02873.txt source-document02250.txt source-document03920.txt source-document03516.txt source-document05608.txt source-document02559.txt source-document05686.txt	source-document02552.txt source-document05393.txt source-document01886.txt source-document01868.txt source-document01496.txt source-document06263.txt source-document00777.txt source-document01833.txt source-document04693.txt source-document00991.txt source-document03674.txt source-document02457.txt source-document06762.txt source-document02811.txt
11	suspicious-document13327.txt	source-document11976.txt source-document10950.txt source-document13628.txt source-document08690.txt source-document08781.txt source-document14392.txt	source-document11055.txt source-document13628.txt source-document12857.txt source-document08941.txt source-document07782.txt source-document10802.txt
9	suspicious-document05387.txt	source-document04848.txt source-document04319.txt source-document06401.txt source-document00060.txt source-document00734.txt	source-document06423.txt source-document04185.txt source-document00439.txt source-document02994.txt
8	suspicious-document02415.txt	source-document00548.txt source-document00006.txt source-document02099.txt source-document02748.txt	source-document03542.txt source-document03601.txt source-document01595.txt source-document02950.txt
8	suspicious-document09248.txt	source-document12172.txt source-document07493.txt source-document11336.txt source-document11685.txt	source-document09681.txt source-document09262.txt source-document11151.txt source-document13475.txt
7	suspicious-document01491.txt	source-document00658.txt source-document03578.txt source-document00128.txt source-document01810.txt	source-document02629.txt source-document00808.txt source-document03364.txt
7	suspicious-document06601.txt	source-document00980.txt source-document03695.txt source-document06880.txt source-document02158.txt	source-document06718.txt source-document04649.txt source-document03267.txt

<b>Fuentes</b>	<b>Nombre Sospechoso</b>	<b>Nombre Fuentes</b>	
7	suspicious-document11262.txt	source-document11331.txt source-document07727.txt source-document13560.txt source-document07615.txt	source-document07715.txt source-document08751.txt source-document09793.txt
6	suspicious-document01203.txt	source-document03559.txt source-document03582.txt source-document04730.txt	source-document02941.txt source-document00739.txt source-document05582.txt
6	suspicious-document05259.txt	source-document00315.txt source-document04563.txt source-document00992.txt	source-document07054.txt source-document01795.txt source-document04173.txt
6	suspicious-document09317.txt	source-document14217.txt source-document10767.txt source-document10006.txt	source-document07755.txt source-document11792.txt source-document09133.txt
6	suspicious-document12515.txt	source-document12637.txt source-document09038.txt source-document13261.txt	source-document08143.txt source-document09983.txt source-document09199.txt
5	suspicious-document01001.txt	source-document05662.txt source-document02021.txt source-document05860.txt	source-document02667.txt source-document05414.txt
5	suspicious-document04061.txt	source-document05023.txt source-document04736.txt source-document06872.txt	source-document01700.txt source-document07069.txt
5	suspicious-document07564.txt	source-document09056.txt source-document09144.txt source-document08032.txt	source-document10523.txt source-document12828.txt
5	suspicious-document10297.txt	source-document08698.txt source-document07906.txt source-document14413.txt	source-document10780.txt source-document07775.txt
5	suspicious-document13596.txt	source-document08029.txt source-document14343.txt source-document12012.txt	source-document07937.txt source-document14040.txt
4	suspicious-document01772.txt	source-document03078.txt source-document01268.txt	source-document03119.txt source-document02906.txt
4	suspicious-document03536.txt	source-document00988.txt source-document01035.txt	source-document04092.txt source-document05120.txt
4	suspicious-document05160.txt	source-document06158.txt source-document06174.txt	source-document04803.txt source-document04049.txt
4	suspicious-document07258.txt	source-document08893.txt source-document10633.txt	source-document09009.txt source-document11540.txt
4	suspicious-document08828.txt	source-document13745.txt source-document09669.txt	source-document07446.txt source-document12367.txt
4	suspicious-document10320.txt	source-document07794.txt source-document12348.txt	source-document12272.txt source-document13734.txt

<b>Fuentes</b>	<b>Nombre Sospechoso</b>	<b>Nombre Fuentes</b>	
4	suspicious-document11966.txt	source-document11740.txt	source-document14084.txt
		source-document12635.txt	source-document09939.txt
4	suspicious-document13588.txt	source-document07844.txt	source-document07937.txt
		source-document09106.txt	source-document14108.txt
3	suspicious-document00256.txt	source-document05840.txt	source-document00191.txt
		source-document05750.txt	
3	suspicious-document01258.txt	source-document02053.txt	source-document03091.txt
		source-document04186.txt	
3	suspicious-document02265.txt	source-document05569.txt	source-document06048.txt
		source-document00063.txt	
3	suspicious-document03320.txt	source-document06637.txt	source-document03258.txt
		source-document03180.txt	
3	suspicious-document04306.txt	source-document05574.txt	source-document03527.txt
		source-document05599.txt	
3	suspicious-document05385.txt	source-document06535.txt	source-document01109.txt
		source-document00971.txt	
3	suspicious-document06694.txt	source-document02444.txt	source-document00159.txt
		source-document06941.txt	
3	suspicious-document07854.txt	source-document09925.txt	source-document07316.txt
		source-document10133.txt	
3	suspicious-document09024.txt	source-document11504.txt	source-document07666.txt
		source-document13049.txt	
3	suspicious-document10044.txt	source-document10826.txt	source-document09514.txt
		source-document09083.txt	
3	suspicious-document11137.txt	source-document10021.txt	source-document13186.txt
		source-document11570.txt	
3	suspicious-document12012.txt	source-document12895.txt	source-document13814.txt
		source-document11497.txt	
3	suspicious-document12876.txt	source-document12079.txt	source-document12450.txt
		source-document08450.txt	
3	suspicious-document13920.txt	source-document11008.txt	source-document11227.txt
		source-document09750.txt	
2	suspicious-document00434.txt	source-document03242.txt	source-document00198.txt
2	suspicious-document01145.txt	source-document02326.txt	source-document03016.txt
2	suspicious-document01806.txt	source-document00617.txt	source-document01788.txt
2	suspicious-document02542.txt	source-document02284.txt	source-document04725.txt
2	suspicious-document03238.txt	source-document05032.txt	source-document02819.txt
2	suspicious-document03980.txt	source-document03720.txt	source-document01446.txt
2	suspicious-document04724.txt	source-document02965.txt	source-document05826.txt
2	suspicious-document05453.txt	source-document02402.txt	source-document04994.txt
2	suspicious-document06100.txt	source-document01700.txt	source-document03049.txt
2	suspicious-document06706.txt	source-document01389.txt	source-document06107.txt
2	suspicious-document07449.txt	source-document12830.txt	source-document13554.txt

Fuentes	Nombre Sospechoso	Nombre Fuentes	
2	suspicious-document08227.txt	source-document11368.txt	source-document10381.txt
2	suspicious-document08870.txt	source-document08713.txt	source-document12337.txt
2	suspicious-document09646.txt	source-document12904.txt	source-document10973.txt
2	suspicious-document10350.txt	source-document10723.txt	source-document07254.txt
2	suspicious-document11082.txt	source-document09962.txt	source-document08388.txt
2	suspicious-document11741.txt	source-document10662.txt	source-document13964.txt
2	suspicious-document12534.txt	source-document12532.txt	source-document07965.txt
2	suspicious-document13370.txt	source-document11311.txt	source-document08003.txt
2	suspicious-document14194.txt	source-document14017.txt	source-document10832.txt
1	suspicious-document00373.txt	source-document01495.txt	
1	suspicious-document00958.txt	source-document00150.txt	
1	suspicious-document01542.txt	source-document00543.txt	
1	suspicious-document02151.txt	source-document01489.txt	
1	suspicious-document02841.txt	source-document05154.txt	
1	suspicious-document03525.txt	source-document03130.txt	
1	suspicious-document04104.txt	source-document06473.txt	
1	suspicious-document04697.txt	source-document01476.txt	
1	suspicious-document05355.txt	source-document00799.txt	
1	suspicious-document05993.txt	source-document00551.txt	
1	suspicious-document06638.txt	source-document06281.txt	
1	suspicious-document07363.txt	source-document13262.txt	
1	suspicious-document08124.txt	source-document14192.txt	
1	suspicious-document08734.txt	source-document07755.txt	
1	suspicious-document09428.txt	source-document12082.txt	
1	suspicious-document10151.txt	source-document10363.txt	
1	suspicious-document10784.txt	source-document11056.txt	
1	suspicious-document11568.txt	source-document12844.txt	
1	suspicious-document12292.txt	source-document09798.txt	
1	suspicious-document13085.txt	source-document12136.txt	
1	suspicious-document14418.txt	source-document08060.txt	

Cuadro D.2: Lista de Documentos Sospechosos y sus respectivos Documentos Fuentes

	MIN	THR	MAX
Cut	0	0.001	1
P-segment	50	1	50

Cuadro D.3: Rango para el corte óptimo entre 0 y 1



ID	P-segment	Cut	Accuracy mean	Accuracy StdDev	Precision mean	Precision StdDev	Recall mean	Recall StdDev	Fmeasure mean	Fmeasure StdDev
Cut1	50	0.020	0.997200	0.004976	0.955628	0.179979	0.676502	0.298889	0.801152	0.186899
Cut2	50	0.041	0.996825	0.005556	0.955628	0.179979	0.647322	0.306781	0.774504	0.206292
Cut3	50	0.062	0.996475	0.005901	0.955628	0.179979	0.611796	0.325165	0.737680	0.236016
Cut4	50	0.083	0.996250	0.006078	0.954444	0.182215	0.568120	0.334802	0.712263	0.242320
Cut5	50	0.104	0.995425	0.006206	0.926596	0.226199	0.383517	0.387385	0.764676	0.195372
Cut6	50	0.125	0.995375	0.006218	0.926596	0.226199	0.378606	0.385523	0.756987	0.199526
Cut7	50	0.146	0.995275	0.006215	0.924823	0.225289	0.364856	0.376963	0.741907	0.204297
Cut8	50	0.167	0.995175	0.006210	0.929078	0.224779	0.349856	0.370719	0.719871	0.216046
Cut9	50	0.188	0.995025	0.006181	0.929078	0.224779	0.333279	0.360473	0.695143	0.218668
Cut10	50	0.209	0.994975	0.006189	0.929078	0.224779	0.329931	0.357921	0.690507	0.217460
Cut11	50	0.230	0.994900	0.006196	0.916667	0.257144	0.311806	0.353494	0.692280	0.223933
Cut12	50	0.251	0.994850	0.006195	0.916667	0.257144	0.306597	0.350052	0.683975	0.224288
Cut13	50	0.272	0.994725	0.006382	0.916667	0.257144	0.295213	0.345922	0.662789	0.233734
Cut14	50	0.293	0.994650	0.006425	0.916667	0.257144	0.287922	0.343780	0.648295	0.242474
Cut15	50	0.314	0.994525	0.006440	0.899225	0.292321	0.268130	0.348517	0.658010	0.252698
Cut16	50	0.335	0.994275	0.006401	0.892473	0.297987	0.183859	0.299457	0.644292	0.230561
Cut17	50	0.356	0.994250	0.006568	0.922222	0.253616	0.181627	0.297699	0.637596	0.234534
Cut18	50	0.377	0.994250	0.006568	0.922222	0.253616	0.181627	0.297699	0.637596	0.234534
Cut19	50	0.398	0.994250	0.006568	0.922222	0.253616	0.181627	0.297699	0.637596	0.234534
Cut20	50	0.419	0.994250	0.006568	0.922222	0.253616	0.181627	0.297699	0.637596	0.234534
Cut21	50	0.440	0.994225	0.006745	0.922222	0.253616	0.181181	0.297891	0.635297	0.239723
Cut22	50	0.461	0.994225	0.006745	0.922222	0.253616	0.181181	0.297891	0.635297	0.239723
Cut23	50	0.482	0.994225	0.006745	0.922222	0.253616	0.181181	0.297891	0.635297	0.239723
Cut24	50	0.503	0.994200	0.006742	0.922222	0.253616	0.178681	0.295192	0.628920	0.238961
Cut25	50	0.524	0.994175	0.006739	0.922222	0.253616	0.175556	0.293090	0.619396	0.242554
Cut26	50	0.545	0.994175	0.006739	0.922222	0.253616	0.175556	0.293090	0.619396	0.242554
Cut27	50	0.566	0.994200	0.006742	0.933333	0.249444	0.175556	0.293090	0.622797	0.242525
Cut28	50	0.587	0.994150	0.006773	0.933333	0.249444	0.172381	0.290948	0.613121	0.245721
Cut29	50	0.608	0.994100	0.006811	0.933333	0.249444	0.168909	0.289866	0.600597	0.255655
Cut30	50	0.629	0.994075	0.006844	0.933333	0.249444	0.167520	0.289874	0.594753	0.262776
Cut31	50	0.650	0.994075	0.006844	0.933333	0.249444	0.167520	0.289874	0.594753	0.262776
Cut32	50	0.671	0.994025	0.006852	0.931034	0.253395	0.159484	0.287687	0.584888	0.273651
Cut33	50	0.692	0.994025	0.006852	0.931034	0.253395	0.159484	0.287687	0.584888	0.273651
Cut34	50	0.713	0.994025	0.006852	0.931034	0.253395	0.159484	0.287687	0.584888	0.273651
Cut35	50	0.734	0.994025	0.006852	0.931034	0.253395	0.159484	0.287687	0.584888	0.273651
Cut36	50	0.755	0.994000	0.006834	0.931034	0.253395	0.155317	0.277725	0.577481	0.264881
Cut37	50	0.776	0.994000	0.006834	0.931034	0.253395	0.155317	0.277725	0.577481	0.264881
Cut38	50	0.797	0.993975	0.006823	0.931034	0.253395	0.151151	0.272473	0.566370	0.261585
Cut39	50	0.818	0.993975	0.006823	0.931034	0.253395	0.151151	0.272473	0.566370	0.261585
Cut40	50	0.839	0.993925	0.006830	0.928571	0.257539	0.143115	0.269659	0.555033	0.271700
Cut41	50	0.860	0.993925	0.006830	0.928571	0.257539	0.143115	0.269659	0.555033	0.271700
Cut42	50	0.881	0.993950	0.006841	0.962963	0.188853	0.143115	0.269659	0.555033	0.271700
Cut43	50	0.902	0.993900	0.006840	0.962963	0.188853	0.137490	0.265425	0.535620	0.275262
Cut44	50	0.923	0.993900	0.006840	0.962963	0.188853	0.137490	0.265425	0.535620	0.275262
Cut45	50	0.944	0.993900	0.006840	0.962963	0.188853	0.137490	0.265425	0.535620	0.275262
Cut46	50	0.965	0.993900	0.006840	0.962963	0.188853	0.137490	0.265425	0.535620	0.275262
Cut47	50	0.986	0.993900	0.006840	0.962963	0.188853	0.137490	0.265425	0.535620	0.275262
<b>Best Cut</b>	50	0.002	0.997525	0.003229	0.956751	0.177823	0.709091	0.280132	0.812790	0.182552

Cuadro D.4: Resultados para el corte óptimo entre 0 y 1

	MIN	THR	MAX
Cut	0	0.000004	0.004
P-segment	50	1	50

Cuadro D.5: Rango para el corte óptimo entre 0 y 0.004

ID	P-segment	Cut	Accuracy mean	Accuracy StdDev	Precision mean	Precision StdDev	Recall mean	Recall StdDev	Fmeasure mean	Fmeasure StdDev
Cut1	50	0.000080	0.994750	0.026806	0.957394	0.176220	0.747037	0.257174	0.814883	0.198139
Cut2	50	0.000164	0.994725	0.026809	0.957394	0.176220	0.744954	0.261161	0.812171	0.203816
Cut3	50	0.000248	0.994725	0.026809	0.957394	0.176220	0.744954	0.261161	0.812171	0.203816
Cut4	50	0.000332	0.997650	0.003127	0.959792	0.164183	0.744954	0.261161	0.816184	0.190811
Cut5	50	0.000416	0.997600	0.003262	0.959792	0.164183	0.740341	0.259912	0.813310	0.189759
Cut6	50	0.000500	0.997600	0.003262	0.959792	0.164183	0.740341	0.259912	0.813310	0.189759
Cut7	50	0.000584	0.997575	0.003266	0.959283	0.165157	0.734091	0.271375	0.815190	0.190239
Cut8	50	0.000668	0.997575	0.003266	0.959283	0.165157	0.734091	0.271375	0.815190	0.190239
Cut9	50	0.000752	0.997550	0.003255	0.959283	0.165157	0.727841	0.270937	0.810917	0.189782
Cut10	50	0.000836	0.997550	0.003255	0.959283	0.165157	0.727841	0.270937	0.810917	0.189782
Cut11	50	0.000920	0.997550	0.003255	0.959283	0.165157	0.727841	0.270937	0.810917	0.189782
Cut12	50	0.001004	0.997550	0.003255	0.959283	0.165157	0.727841	0.270937	0.810917	0.189782
Cut13	50	0.001088	0.997550	0.003255	0.959283	0.165157	0.727841	0.270937	0.810917	0.189782
Cut14	50	0.001172	0.997550	0.003255	0.959283	0.165157	0.727841	0.270937	0.810917	0.189782
Cut15	50	0.001256	0.997550	0.003255	0.959283	0.165157	0.727841	0.270937	0.810917	0.189782
Cut16	50	0.001340	0.997525	0.003275	0.959283	0.165157	0.725758	0.272038	0.809207	0.190472
Cut17	50	0.001424	0.997500	0.003263	0.959283	0.165157	0.721591	0.270353	0.806643	0.189229
Cut18	50	0.001508	0.997500	0.003263	0.959283	0.165157	0.721591	0.270353	0.806643	0.189229
Cut19	50	0.001592	0.997500	0.003263	0.959283	0.165157	0.721591	0.270353	0.806643	0.189229
Cut20	50	0.001676	0.997525	0.003229	0.959916	0.162317	0.721591	0.270353	0.807498	0.187229
Cut21	50	0.001760	0.997525	0.003229	0.956751	0.177823	0.709091	0.280132	0.812790	0.182552
Cut22	50	0.001844	0.997525	0.003229	0.956751	0.177823	0.709091	0.280132	0.812790	0.182552
Cut23	50	0.001928	0.997525	0.003229	0.956751	0.177823	0.709091	0.280132	0.812790	0.182552
Cut24	50	0.002012	0.997525	0.003229	0.956751	0.177823	0.709091	0.280132	0.812790	0.182552
Cut25	50	0.002096	0.997525	0.003229	0.956751	0.177823	0.709091	0.280132	0.812790	0.182552
Cut26	50	0.002180	0.997525	0.003229	0.956751	0.177823	0.709091	0.280132	0.812790	0.182552
Cut27	50	0.002264	0.997525	0.003229	0.956751	0.177823	0.709091	0.280132	0.812790	0.182552
Cut28	50	0.002348	0.997525	0.003229	0.956751	0.177823	0.709091	0.280132	0.812790	0.182552
Cut29	50	0.002432	0.997525	0.003229	0.956751	0.177823	0.709091	0.280132	0.812790	0.182552
Cut30	50	0.002516	0.997525	0.003229	0.956751	0.177823	0.709091	0.280132	0.812790	0.182552
Cut31	50	0.002600	0.997525	0.003229	0.956751	0.177823	0.709091	0.280132	0.812790	0.182552
Cut32	50	0.002684	0.997525	0.003229	0.956751	0.177823	0.709091	0.280132	0.812790	0.182552
Cut33	50	0.002768	0.997525	0.003229	0.956751	0.177823	0.709091	0.280132	0.812790	0.182552
Cut34	50	0.002852	0.997525	0.003229	0.956751	0.177823	0.709091	0.280132	0.812790	0.182552
Cut35	50	0.002936	0.997525	0.003229	0.956751	0.177823	0.709091	0.280132	0.812790	0.182552
Cut36	50	0.003020	0.997525	0.003229	0.956751	0.177823	0.709091	0.280132	0.812790	0.182552
Cut37	50	0.003104	0.997500	0.003217	0.956197	0.178892	0.696591	0.289042	0.810327	0.182474
Cut38	50	0.003188	0.997500	0.003217	0.956197	0.178892	0.696591	0.289042	0.810327	0.182474
Cut39	50	0.003272	0.997475	0.003373	0.956197	0.178892	0.696144	0.289207	0.809955	0.182614
Cut40	50	0.003356	0.997475	0.003373	0.956197	0.178892	0.696144	0.289207	0.809955	0.182614
Cut41	50	0.003440	0.997475	0.003373	0.956197	0.178892	0.696144	0.289207	0.809955	0.182614
Cut42	50	0.003524	0.997475	0.003373	0.956197	0.178892	0.696144	0.289207	0.809955	0.182614
Cut43	50	0.003608	0.997475	0.003373	0.956197	0.178892	0.696144	0.289207	0.809955	0.182614
Cut44	50	0.003692	0.997475	0.003373	0.956197	0.178892	0.696144	0.289207	0.809955	0.182614
Cut45	50	0.003776	0.997475	0.003373	0.956197	0.178892	0.696144	0.289207	0.809955	0.182614
Cut46	50	0.003860	0.997475	0.003373	0.956197	0.178892	0.696144	0.289207	0.809955	0.182614
Cut47	50	0.003944	0.997450	0.003535	0.956197	0.178892	0.695698	0.289427	0.809565	0.182821
<b>Best Cut</b>	50	0.000320	0.997650	0.003127	0.959792	0.164183	0.744954	0.261161	0.816184	0.190811

Cuadro D.6: Resultados para el corte óptimo entre 0 y 0.004

	MIN	THR	MAX
Cut	0.00022	0.0000004	0.00042
P-segment	50	1	50

Cuadro D.7: Rango para el corte óptimo entre 0.00022 y 0.00042

ID	P-segment	Cut	Accuracy mean	Accuracy StdDev	Precision mean	Precision StdDev	Recall mean	Recall StdDev	Fmeasure mean	Fmeasure StdDev
Cut1	50	0.000224	0.994725	0.026809	0.957394	0.176220	0.744954	0.261161	0.812171	0.203816
Cut2	50	0.000228	0.994725	0.026809	0.957394	0.176220	0.744954	0.261161	0.812171	0.203816
Cut3	50	0.000232	0.994725	0.026809	0.957394	0.176220	0.744954	0.261161	0.812171	0.203816
Cut4	50	0.000237	0.994725	0.026809	0.957394	0.176220	0.744954	0.261161	0.812171	0.203816
Cut5	50	0.000241	0.994725	0.026809	0.957394	0.176220	0.744954	0.261161	0.812171	0.203816
Cut6	50	0.000245	0.994725	0.026809	0.957394	0.176220	0.744954	0.261161	0.812171	0.203816
Cut7	50	0.000249	0.994725	0.026809	0.957394	0.176220	0.744954	0.261161	0.812171	0.203816
Cut8	50	0.000253	0.994725	0.026809	0.957394	0.176220	0.744954	0.261161	0.812171	0.203816
Cut9	50	0.000258	0.994725	0.026809	0.957394	0.176220	0.744954	0.261161	0.812171	0.203816
Cut10	50	0.000262	0.994725	0.026809	0.957394	0.176220	0.744954	0.261161	0.812171	0.203816
Cut11	50	0.000266	0.994725	0.026809	0.957394	0.176220	0.744954	0.261161	0.812171	0.203816
Cut12	50	0.000270	0.994725	0.026809	0.957394	0.176220	0.744954	0.261161	0.812171	0.203816
Cut13	50	0.000274	0.994725	0.026809	0.957394	0.176220	0.744954	0.261161	0.812171	0.203816
Cut14	50	0.000279	0.994725	0.026809	0.957394	0.176220	0.744954	0.261161	0.812171	0.203816
Cut15	50	0.000283	0.994725	0.026809	0.957394	0.176220	0.744954	0.261161	0.812171	0.203816
Cut16	50	0.000287	0.994725	0.026809	0.957394	0.176220	0.744954	0.261161	0.812171	0.203816
Cut17	50	0.000291	0.994725	0.026809	0.957394	0.176220	0.744954	0.261161	0.812171	0.203816
Cut18	50	0.000295	0.994725	0.026809	0.957394	0.176220	0.744954	0.261161	0.812171	0.203816
Cut19	50	0.000300	0.994725	0.026809	0.957394	0.176220	0.744954	0.261161	0.812171	0.203816
Cut20	50	0.000304	0.994725	0.026809	0.957394	0.176220	0.744954	0.261161	0.812171	0.203816
Cut21	50	0.000308	0.994725	0.026809	0.957394	0.176220	0.744954	0.261161	0.812171	0.203816
Cut22	50	0.000312	0.994725	0.026809	0.957394	0.176220	0.744954	0.261161	0.812171	0.203816
Cut23	50	0.000316	0.997650	0.003127	0.959792	0.164183	0.744954	0.261161	0.816184	0.190811
Cut24	50	0.000321	0.997650	0.003127	0.959792	0.164183	0.744954	0.261161	0.816184	0.190811
Cut25	50	0.000325	0.997650	0.003127	0.959792	0.164183	0.744954	0.261161	0.816184	0.190811
Cut26	50	0.000329	0.997650	0.003127	0.959792	0.164183	0.744954	0.261161	0.816184	0.190811
Cut27	50	0.000333	0.997650	0.003127	0.959792	0.164183	0.744954	0.261161	0.816184	0.190811
Cut28	50	0.000337	0.997650	0.003127	0.959792	0.164183	0.744954	0.261161	0.816184	0.190811
Cut29	50	0.000342	0.997650	0.003127	0.959792	0.164183	0.744954	0.261161	0.816184	0.190811
Cut30	50	0.000346	0.997650	0.003127	0.959792	0.164183	0.744954	0.261161	0.816184	0.190811
Cut31	50	0.000350	0.997650	0.003127	0.959792	0.164183	0.744954	0.261161	0.816184	0.190811
Cut32	50	0.000354	0.997650	0.003127	0.959792	0.164183	0.744954	0.261161	0.816184	0.190811
Cut33	50	0.000358	0.997650	0.003127	0.959792	0.164183	0.744954	0.261161	0.816184	0.190811
Cut34	50	0.000363	0.997650	0.003127	0.959792	0.164183	0.744954	0.261161	0.816184	0.190811
Cut35	50	0.000367	0.997650	0.003127	0.959792	0.164183	0.744954	0.261161	0.816184	0.190811
Cut36	50	0.000371	0.997650	0.003127	0.959792	0.164183	0.744954	0.261161	0.816184	0.190811
Cut37	50	0.000375	0.997650	0.003127	0.959792	0.164183	0.744954	0.261161	0.816184	0.190811
Cut38	50	0.000379	0.997625	0.003116	0.959792	0.164183	0.740787	0.259714	0.813653	0.189678
Cut39	50	0.000384	0.997625	0.003116	0.959792	0.164183	0.740787	0.259714	0.813653	0.189678
Cut40	50	0.000388	0.997625	0.003116	0.959792	0.164183	0.740787	0.259714	0.813653	0.189678
Cut41	50	0.000392	0.997625	0.003116	0.959792	0.164183	0.740787	0.259714	0.813653	0.189678
Cut42	50	0.000396	0.997600	0.003262	0.959792	0.164183	0.740341	0.259912	0.813310	0.189759
Cut43	50	0.000400	0.997600	0.003262	0.959792	0.164183	0.740341	0.259912	0.813310	0.189759
Cut44	50	0.000405	0.997600	0.003262	0.959792	0.164183	0.740341	0.259912	0.813310	0.189759
Cut45	50	0.000409	0.997600	0.003262	0.959792	0.164183	0.740341	0.259912	0.813310	0.189759
Cut46	50	0.000413	0.997600	0.003262	0.959792	0.164183	0.740341	0.259912	0.813310	0.189759
Cut47	50	0.000417	0.997600	0.003262	0.959792	0.164183	0.740341	0.259912	0.813310	0.189759
<b>Best Cut</b>	50	0.000316	0.997650	0.003127	0.959792	0.164183	0.744954	0.261161	0.816184	0.190811

Cuadro D.8: Resultados para el corte óptimo entre 0.00022 y 0.00042

	MIN	THR	MAX
Cut	0.0003122	0.00000008	0.0003922
P-segment	50	1	50

Cuadro D.9: Rango para el corte óptimo entre 0.0003122 y 0.0003922

ID	P-segment	Cut	Accuracy mean	Accuracy StdDev	Precision mean	Precision StdDev	Recall mean	Recall StdDev	Fmeasure mean	Fmeasure StdDev
Cut1	50	0.000314	0.994725	0.026809	0.957394	0.176220	0.744954	0.261161	0.812171	0.203816
Cut2	50	0.000315	0.994725	0.026809	0.957394	0.176220	0.744954	0.261161	0.812171	0.203816
Cut3	50	0.000317	0.997650	0.003127	0.959792	0.164183	0.744954	0.261161	0.816184	0.190811
Cut4	50	0.000319	0.997650	0.003127	0.959792	0.164183	0.744954	0.261161	0.816184	0.190811
Cut5	50	0.000321	0.997650	0.003127	0.959792	0.164183	0.744954	0.261161	0.816184	0.190811
Cut6	50	0.000322	0.997650	0.003127	0.959792	0.164183	0.744954	0.261161	0.816184	0.190811
Cut7	50	0.000324	0.997650	0.003127	0.959792	0.164183	0.744954	0.261161	0.816184	0.190811
Cut8	50	0.000326	0.997650	0.003127	0.959792	0.164183	0.744954	0.261161	0.816184	0.190811
Cut9	50	0.000327	0.997650	0.003127	0.959792	0.164183	0.744954	0.261161	0.816184	0.190811
Cut10	50	0.000329	0.997650	0.003127	0.959792	0.164183	0.744954	0.261161	0.816184	0.190811
Cut11	50	0.000331	0.997650	0.003127	0.959792	0.164183	0.744954	0.261161	0.816184	0.190811
Cut12	50	0.000332	0.997650	0.003127	0.959792	0.164183	0.744954	0.261161	0.816184	0.190811
Cut13	50	0.000334	0.997650	0.003127	0.959792	0.164183	0.744954	0.261161	0.816184	0.190811
Cut14	50	0.000336	0.997650	0.003127	0.959792	0.164183	0.744954	0.261161	0.816184	0.190811
Cut15	50	0.000337	0.997650	0.003127	0.959792	0.164183	0.744954	0.261161	0.816184	0.190811
Cut16	50	0.000339	0.997650	0.003127	0.959792	0.164183	0.744954	0.261161	0.816184	0.190811
Cut17	50	0.000341	0.997650	0.003127	0.959792	0.164183	0.744954	0.261161	0.816184	0.190811
Cut18	50	0.000342	0.997650	0.003127	0.959792	0.164183	0.744954	0.261161	0.816184	0.190811
Cut19	50	0.000344	0.997650	0.003127	0.959792	0.164183	0.744954	0.261161	0.816184	0.190811
Cut20	50	0.000346	0.997650	0.003127	0.959792	0.164183	0.744954	0.261161	0.816184	0.190811
Cut21	50	0.000347	0.997650	0.003127	0.959792	0.164183	0.744954	0.261161	0.816184	0.190811
Cut22	50	0.000349	0.997650	0.003127	0.959792	0.164183	0.744954	0.261161	0.816184	0.190811
Cut23	50	0.000351	0.997650	0.003127	0.959792	0.164183	0.744954	0.261161	0.816184	0.190811
Cut24	50	0.000352	0.997650	0.003127	0.959792	0.164183	0.744954	0.261161	0.816184	0.190811
Cut25	50	0.000354	0.997650	0.003127	0.959792	0.164183	0.744954	0.261161	0.816184	0.190811
Cut26	50	0.000356	0.997650	0.003127	0.959792	0.164183	0.744954	0.261161	0.816184	0.190811
Cut27	50	0.000357	0.997650	0.003127	0.959792	0.164183	0.744954	0.261161	0.816184	0.190811
Cut28	50	0.000359	0.997650	0.003127	0.959792	0.164183	0.744954	0.261161	0.816184	0.190811
Cut29	50	0.000361	0.997650	0.003127	0.959792	0.164183	0.744954	0.261161	0.816184	0.190811
Cut30	50	0.000363	0.997650	0.003127	0.959792	0.164183	0.744954	0.261161	0.816184	0.190811
Cut31	50	0.000364	0.997650	0.003127	0.959792	0.164183	0.744954	0.261161	0.816184	0.190811
Cut32	50	0.000366	0.997650	0.003127	0.959792	0.164183	0.744954	0.261161	0.816184	0.190811
Cut33	50	0.000368	0.997650	0.003127	0.959792	0.164183	0.744954	0.261161	0.816184	0.190811
Cut34	50	0.000369	0.997650	0.003127	0.959792	0.164183	0.744954	0.261161	0.816184	0.190811
Cut35	50	0.000371	0.997650	0.003127	0.959792	0.164183	0.744954	0.261161	0.816184	0.190811
Cut36	50	0.000373	0.997650	0.003127	0.959792	0.164183	0.744954	0.261161	0.816184	0.190811
Cut37	50	0.000374	0.997650	0.003127	0.959792	0.164183	0.744954	0.261161	0.816184	0.190811
Cut38	50	0.000376	0.997650	0.003127	0.959792	0.164183	0.744954	0.261161	0.816184	0.190811
Cut39	50	0.000378	0.997650	0.003127	0.959792	0.164183	0.744954	0.261161	0.816184	0.190811
Cut40	50	0.000379	0.997625	0.003116	0.959792	0.164183	0.740787	0.259714	0.813653	0.189678
Cut41	50	0.000381	0.997625	0.003116	0.959792	0.164183	0.740787	0.259714	0.813653	0.189678
Cut42	50	0.000383	0.997625	0.003116	0.959792	0.164183	0.740787	0.259714	0.813653	0.189678
Cut43	50	0.000384	0.997625	0.003116	0.959792	0.164183	0.740787	0.259714	0.813653	0.189678
Cut44	50	0.000386	0.997625	0.003116	0.959792	0.164183	0.740787	0.259714	0.813653	0.189678
Cut45	50	0.000388	0.997625	0.003116	0.959792	0.164183	0.740787	0.259714	0.813653	0.189678
Cut46	50	0.000389	0.997625	0.003116	0.959792	0.164183	0.740787	0.259714	0.813653	0.189678
Cut47	50	0.000391	0.997625	0.003116	0.959792	0.164183	0.740787	0.259714	0.813653	0.189678
<b>Best Cut</b>	50	0.000316	0.997650	0.003127	0.959792	0.164183	0.744954	0.261161	0.816184	0.190811

Cuadro D.10: Resultados para el corte óptimo entre 0.0003122 y 0.0003922

	MIN	THR	MAX
Cut	0.00031628	0.1	0.00031628
P-segment	2	5	1000

Cuadro D.11: Rango para el p-segmento óptimo entre 2 y 1000

ID	P-segment	Cut	Accuracy mean	Accuracy StdDev	Precision mean	Precision StdDev	Recall mean	Recall StdDev	Fmeasure mean	Fmeasure StdDev
P-segment1	22	0.000316	0.997150	0.004688	0.959792	0.164183	0.701531	0.272838	0.782555	0.201435
P-segment2	47	0.000316	0.997600	0.003292	0.959792	0.164183	0.742424	0.262639	0.814154	0.191615
P-segment3	72	0.000316	0.997775	0.002898	0.960833	0.162577	0.755912	0.256900	0.824750	0.183922
P-segment4	97	0.000316	0.997800	0.002891	0.960833	0.162577	0.757698	0.257100	0.825886	0.184248
P-segment5	122	0.000316	0.997875	0.002537	0.960833	0.162577	0.759037	0.256964	0.826791	0.184336
P-segment6	147	0.000316	0.997900	0.002508	0.960833	0.162577	0.760823	0.256149	0.828133	0.183992
P-segment7	172	0.000316	0.997975	0.002439	0.960833	0.162577	0.766552	0.254098	0.832315	0.183181
P-segment8	197	0.000316	0.998000	0.002429	0.960833	0.162577	0.767688	0.254529	0.832978	0.183520
P-segment9	222	0.000316	0.998000	0.002429	0.960833	0.162577	0.767688	0.254529	0.832978	0.183520
P-segment10	247	0.000316	0.998025	0.002334	0.960833	0.162577	0.768134	0.254592	0.833256	0.183607
P-segment11	272	0.000316	0.998050	0.002345	0.960833	0.162577	0.771259	0.255881	0.835064	0.184535
P-segment12	297	0.000316	0.998050	0.002345	0.960833	0.162577	0.771259	0.255881	0.835064	0.184535
P-segment13	322	0.000316	0.998075	0.002312	0.960833	0.162577	0.773759	0.253204	0.837324	0.182373
P-segment14	347	0.000316	0.998100	0.002234	0.960833	0.162577	0.774206	0.253320	0.837592	0.182482
P-segment15	372	0.000316	0.998125	0.002244	0.960833	0.162577	0.780456	0.252644	0.841811	0.182335
P-segment16	397	0.000316	0.998125	0.002244	0.960833	0.162577	0.780456	0.252644	0.841811	0.182335
P-segment17	422	0.000316	0.998125	0.002244	0.960833	0.162577	0.780456	0.252644	0.841811	0.182335
P-segment18	447	0.000316	0.998125	0.002244	0.960833	0.162577	0.780456	0.252644	0.841811	0.182335
P-segment19	472	0.000316	0.998150	0.002186	0.960833	0.162577	0.780902	0.252811	0.842068	0.182464
P-segment20	497	0.000316	0.998150	0.002186	0.960833	0.162577	0.780902	0.252811	0.842068	0.182464
P-segment21	522	0.000316	0.998150	0.002186	0.960833	0.162577	0.780902	0.252811	0.842068	0.182464
P-segment22	547	0.000316	0.998125	0.002244	0.960353	0.162517	0.780902	0.252811	0.841847	0.182352
P-segment23	572	0.000316	0.998125	0.002244	0.960353	0.162517	0.780902	0.252811	0.841847	0.182352
P-segment24	597	0.000316	0.998125	0.002244	0.960353	0.162517	0.780902	0.252811	0.841847	0.182352
P-segment25	622	0.000316	0.998125	0.002244	0.960353	0.162517	0.780902	0.252811	0.841847	0.182352
P-segment26	647	0.000316	0.998125	0.002244	0.960353	0.162517	0.780902	0.252811	0.841847	0.182352
P-segment27	672	0.000316	0.998125	0.002244	0.960353	0.162517	0.780902	0.252811	0.841847	0.182352
P-segment28	697	0.000316	0.998125	0.002244	0.960353	0.162517	0.780902	0.252811	0.841847	0.182352
P-segment29	722	0.000316	0.998125	0.002244	0.960353	0.162517	0.780902	0.252811	0.841847	0.182352
P-segment30	747	0.000316	0.998125	0.002244	0.960353	0.162517	0.780902	0.252811	0.841847	0.182352
P-segment31	772	0.000316	0.998125	0.002244	0.960353	0.162517	0.780902	0.252811	0.841847	0.182352
P-segment32	797	0.000316	0.998125	0.002244	0.960353	0.162517	0.780902	0.252811	0.841847	0.182352
P-segment33	822	0.000316	0.998100	0.002256	0.957853	0.163424	0.780902	0.252811	0.840722	0.182332
P-segment34	847	0.000316	0.998100	0.002256	0.957853	0.163424	0.780902	0.252811	0.840722	0.182332
P-segment35	872	0.000316	0.998100	0.002256	0.957853	0.163424	0.780902	0.252811	0.840722	0.182332
P-segment36	897	0.000316	0.998100	0.002256	0.957853	0.163424	0.780902	0.252811	0.840722	0.182332
P-segment37	922	0.000316	0.998100	0.002256	0.957853	0.163424	0.780902	0.252811	0.840722	0.182332
P-segment38	947	0.000316	0.998100	0.002256	0.957853	0.163424	0.780902	0.252811	0.840722	0.182332
P-segment39	972	0.000316	0.998100	0.002256	0.957853	0.163424	0.780902	0.252811	0.840722	0.182332
P-segment40	997	0.000316	0.998100	0.002256	0.957853	0.163424	0.780902	0.252811	0.840722	0.182332
<b>Best P-segment</b>	472	0.000316	0.998150	0.002186	0.960833	0.162577	0.780902	0.252811	0.842068	0.182464

Cuadro D.12: Resultados para el p-segmento óptimo entre 2 y 1000



	MIN	THR	MAX
Cut	0.00031628	0.1	0.00031628
P-segment	400	1	600

Cuadro D.13: Rango para el p-segmento óptimo entre 400 y 600

ID	P-segment	Cut	Accuracy mean	Accuracy StdDev	Precision mean	Precision StdDev	Recall mean	Recall StdDev	Fmeasure mean	Fmeasure StdDev
P-segment1	404	0.000316	0.998125	0.002244	0.960833	0.162577	0.780456	0.252644	0.841811	0.182335
P-segment2	409	0.000316	0.998125	0.002244	0.960833	0.162577	0.780456	0.252644	0.841811	0.182335
P-segment3	414	0.000316	0.998125	0.002244	0.960833	0.162577	0.780456	0.252644	0.841811	0.182335
P-segment4	419	0.000316	0.998125	0.002244	0.960833	0.162577	0.780456	0.252644	0.841811	0.182335
P-segment5	424	0.000316	0.998125	0.002244	0.960833	0.162577	0.780456	0.252644	0.841811	0.182335
P-segment6	429	0.000316	0.998125	0.002244	0.960833	0.162577	0.780456	0.252644	0.841811	0.182335
P-segment7	434	0.000316	0.998125	0.002244	0.960833	0.162577	0.780456	0.252644	0.841811	0.182335
P-segment8	439	0.000316	0.998125	0.002244	0.960833	0.162577	0.780456	0.252644	0.841811	0.182335
P-segment9	444	0.000316	0.998125	0.002244	0.960833	0.162577	0.780456	0.252644	0.841811	0.182335
P-segment10	449	0.000316	0.998125	0.002244	0.960833	0.162577	0.780456	0.252644	0.841811	0.182335
P-segment11	454	0.000316	0.998125	0.002244	0.960833	0.162577	0.780456	0.252644	0.841811	0.182335
P-segment12	459	0.000316	0.998125	0.002244	0.960833	0.162577	0.780456	0.252644	0.841811	0.182335
P-segment13	464	0.000316	0.998125	0.002244	0.960833	0.162577	0.780456	0.252644	0.841811	0.182335
P-segment14	469	0.000316	0.998125	0.002244	0.960833	0.162577	0.780456	0.252644	0.841811	0.182335
P-segment15	474	0.000316	0.998150	0.002186	0.960833	0.162577	0.780902	0.252811	0.842068	0.182464
P-segment16	479	0.000316	0.998150	0.002186	0.960833	0.162577	0.780902	0.252811	0.842068	0.182464
P-segment17	484	0.000316	0.998150	0.002186	0.960833	0.162577	0.780902	0.252811	0.842068	0.182464
P-segment18	489	0.000316	0.998150	0.002186	0.960833	0.162577	0.780902	0.252811	0.842068	0.182464
P-segment19	494	0.000316	0.998150	0.002186	0.960833	0.162577	0.780902	0.252811	0.842068	0.182464
P-segment20	499	0.000316	0.998150	0.002186	0.960833	0.162577	0.780902	0.252811	0.842068	0.182464
P-segment21	504	0.000316	0.998150	0.002186	0.960833	0.162577	0.780902	0.252811	0.842068	0.182464
P-segment22	509	0.000316	0.998150	0.002186	0.960833	0.162577	0.780902	0.252811	0.842068	0.182464
P-segment23	514	0.000316	0.998150	0.002186	0.960833	0.162577	0.780902	0.252811	0.842068	0.182464
P-segment24	519	0.000316	0.998150	0.002186	0.960833	0.162577	0.780902	0.252811	0.842068	0.182464
P-segment25	524	0.000316	0.998150	0.002186	0.960833	0.162577	0.780902	0.252811	0.842068	0.182464
P-segment26	529	0.000316	0.998125	0.002244	0.960353	0.162517	0.780902	0.252811	0.841847	0.182352
P-segment27	534	0.000316	0.998125	0.002244	0.960353	0.162517	0.780902	0.252811	0.841847	0.182352
P-segment28	539	0.000316	0.998125	0.002244	0.960353	0.162517	0.780902	0.252811	0.841847	0.182352
P-segment29	544	0.000316	0.998125	0.002244	0.960353	0.162517	0.780902	0.252811	0.841847	0.182352
P-segment30	549	0.000316	0.998125	0.002244	0.960353	0.162517	0.780902	0.252811	0.841847	0.182352
P-segment31	554	0.000316	0.998125	0.002244	0.960353	0.162517	0.780902	0.252811	0.841847	0.182352
P-segment32	559	0.000316	0.998125	0.002244	0.960353	0.162517	0.780902	0.252811	0.841847	0.182352
P-segment33	564	0.000316	0.998125	0.002244	0.960353	0.162517	0.780902	0.252811	0.841847	0.182352
P-segment34	569	0.000316	0.998125	0.002244	0.960353	0.162517	0.780902	0.252811	0.841847	0.182352
P-segment35	574	0.000316	0.998125	0.002244	0.960353	0.162517	0.780902	0.252811	0.841847	0.182352
P-segment36	579	0.000316	0.998125	0.002244	0.960353	0.162517	0.780902	0.252811	0.841847	0.182352
P-segment37	584	0.000316	0.998125	0.002244	0.960353	0.162517	0.780902	0.252811	0.841847	0.182352
P-segment38	589	0.000316	0.998125	0.002244	0.960353	0.162517	0.780902	0.252811	0.841847	0.182352
P-segment39	594	0.000316	0.998125	0.002244	0.960353	0.162517	0.780902	0.252811	0.841847	0.182352
P-segment40	599	0.000316	0.998125	0.002244	0.960353	0.162517	0.780902	0.252811	0.841847	0.182352
<b>Best P-segment</b>	471	0.000316	0.998150	0.002186	0.960833	0.162577	0.780902	0.252811	0.842068	0.182464

Cuadro D.14: Resultados para el p-segmento óptimo entre 400 y 600

Suspicious Document	Accuracy	Precision	Recall	Fmeasure
suspicious-document04112	0.994000	1.000000	0.892857	0.943396
suspicious-document13327	0.998000	1.000000	0.909091	0.952381
suspicious-document05387	0.994000	1.000000	0.666667	0.800000
suspicious-document02415	0.996000	1.000000	0.750000	0.857143
suspicious-document09248	0.990000	1.000000	0.375000	0.545455
suspicious-document01491	0.996000	1.000000	0.714286	0.833333
suspicious-document06601	0.998000	1.000000	0.857143	0.923077
suspicious-document11262	0.998000	1.000000	0.857143	0.923077
suspicious-document01203	0.992000	1.000000	0.333333	0.500000
suspicious-document05259	0.998000	1.000000	0.833333	0.909091
suspicious-document09317	0.996000	1.000000	0.666667	0.800000
suspicious-document12515	0.998000	1.000000	0.833333	0.909091
suspicious-document01001	0.998000	1.000000	0.800000	0.888889
suspicious-document04061	1.000000	1.000000	1.000000	1.000000
suspicious-document07564	0.998000	1.000000	0.800000	0.888889
suspicious-document10297	0.996000	1.000000	0.600000	0.750000
suspicious-document13596	0.998000	0.833333	1.000000	0.909091
suspicious-document01772	0.998000	1.000000	0.750000	0.857143
suspicious-document03536	1.000000	1.000000	1.000000	1.000000
suspicious-document05160	1.000000	1.000000	1.000000	1.000000
suspicious-document07258	0.994000	1.000000	0.250000	0.400000
suspicious-document08828	0.994000	1.000000	0.250000	0.400000
suspicious-document10320	0.996000	1.000000	0.500000	0.666667
suspicious-document11966	0.996000	0.750000	0.750000	0.750000
suspicious-document13588	0.996000	0.750000	0.750000	0.750000
suspicious-document00256	1.000000	1.000000	1.000000	1.000000
suspicious-document01258	1.000000	1.000000	1.000000	1.000000
suspicious-document02265	1.000000	1.000000	1.000000	1.000000
suspicious-document03320	1.000000	1.000000	1.000000	1.000000
suspicious-document04306	0.996000	1.000000	0.333333	0.500000
suspicious-document05385	0.998000	1.000000	0.666667	0.800000
suspicious-document06694	0.998000	1.000000	0.666667	0.800000
suspicious-document07854	0.996000	1.000000	0.333333	0.500000
suspicious-document09024	0.998000	1.000000	0.666667	0.800000
suspicious-document10044	1.000000	1.000000	1.000000	1.000000
suspicious-document11137	1.000000	1.000000	1.000000	1.000000
suspicious-document12012	1.000000	1.000000	1.000000	1.000000
suspicious-document12876	1.000000	1.000000	1.000000	1.000000
suspicious-document13920	0.998000	1.000000	0.666667	0.800000
suspicious-document00434	1.000000	1.000000	1.000000	1.000000
suspicious-document01145	1.000000	1.000000	1.000000	1.000000
suspicious-document01806	1.000000	1.000000	1.000000	1.000000

Suspicious Document	Accuracy	Precision	Recall	Fmeasure
suspicious-document02542	0.998000	1.000000	0.500000	0.666667
suspicious-document03238	1.000000	1.000000	1.000000	1.000000
suspicious-document03980	0.998000	1.000000	0.500000	0.666667
suspicious-document04724	1.000000	1.000000	1.000000	1.000000
suspicious-document05453	0.998000	1.000000	0.500000	0.666667
suspicious-document06100	0.998000	1.000000	0.500000	0.666667
suspicious-document06706	0.998000	1.000000	0.500000	0.666667
suspicious-document07449	0.998000	1.000000	0.500000	0.666667
suspicious-document08227	0.998000	1.000000	0.500000	0.666667
suspicious-document08870	0.998000	1.000000	0.500000	0.666667
suspicious-document09646	0.998000	1.000000	0.500000	0.666667
suspicious-document10350	0.998000	1.000000	0.500000	0.666667
suspicious-document11082	1.000000	1.000000	1.000000	1.000000
suspicious-document11741	0.998000	1.000000	0.500000	0.666667
suspicious-document12534	0.998000	1.000000	0.500000	0.666667
suspicious-document13370	0.998000	1.000000	0.500000	0.666667
suspicious-document14194	0.998000	1.000000	0.500000	0.666667
suspicious-document00373	1.000000	1.000000	1.000000	1.000000
suspicious-document00958	1.000000	1.000000	1.000000	1.000000
suspicious-document01542	1.000000	1.000000	1.000000	1.000000
suspicious-document02151	0.992000	0.200000	1.000000	0.333333
suspicious-document02841	1.000000	1.000000	1.000000	1.000000
suspicious-document03525	0.996000	0.000000	0.000000	NaN
suspicious-document04104	1.000000	1.000000	1.000000	1.000000
suspicious-document04697	1.000000	1.000000	1.000000	1.000000
suspicious-document05355	1.000000	1.000000	1.000000	1.000000
suspicious-document05993	1.000000	1.000000	1.000000	1.000000
suspicious-document06638	1.000000	1.000000	1.000000	1.000000
suspicious-document07363	1.000000	1.000000	1.000000	1.000000
suspicious-document08124	1.000000	1.000000	1.000000	1.000000
suspicious-document08734	1.000000	1.000000	1.000000	1.000000
suspicious-document09428	0.996000	0.333333	1.000000	0.500000
suspicious-document10151	1.000000	1.000000	1.000000	1.000000
suspicious-document10784	1.000000	1.000000	1.000000	1.000000
suspicious-document11568	1.000000	1.000000	1.000000	1.000000
suspicious-document12292	1.000000	1.000000	1.000000	1.000000
suspicious-document13085	1.000000	1.000000	1.000000	1.000000
suspicious-document14418	1.000000	1.000000	1.000000	1.000000
Mean	0.998150	0.960833	0.780902	0.842068
StdDev	0.002186	0.162577	0.252811	0.182464

Cuadro D.15: Indicadores para el  $corte = 3,5 \cdot 10^{-4}$  y  $p-segmento = 500$