Actas de las Jenui, vol. 1. 2016.
Páginas: 185-192

Estrategias para programar la detección de plagios en actividades basadas en texto

Juan Ramón Rico-Juan
Departamento de Lenguajes y Sistemas Informáticos
Universidad de Alicante
03080 Alicante
juanramonrico@ua.es

Antonio Javier Gallego
Departamento de Lenguajes y Sistemas Informáticos
Universidad de Alicante
03080 Alicante
jgallego@dlsi.ua.es

José María García Avilés
Unidad de Biblioteca Universitaria
Universidad de Alicante
03080 Alicante
garcia.aviles@ua.es

Resumen

La detección de plagios en los trabajos entregados por los alumnos es un problema que ha existido tradicionalmente cuando se entregaban en formato papel pero que en los últimos años se ha incrementado debido a la gran cantidad de información que existe en Internet, a la facilidad para encontrarla usando buscadores y a la entrega electrónica de los trabajos o actividades (ciberplagio). Incluso existen plataformas en Internet que estructuran y ofrecen gratuitamente los trabajos para que se puedan descargar.

En este artículo se proponen varias estrategias orientadas a implementar un programa para uso personal que detecte uno de los tipos de plagio más extendidos actualmente como es copiar y pegar fragmentos de textos de Internet. Estas propuestas se estudian para la detección de plagio en trabajos de diferente índole, incluyendo memorias, diapositivas y páginas web. El sistema devuelve un índice de coincidencia por entrega, de esta forma el profesor puede identificar claramente las copias y centrar su esfuerzos en revisar solamente el contenido de las tareas originales.

Abstract

Detection of plagiarism in students' homework is a problem that already existed when they were submited in paper form. In recent years, it has increased due to the large amount of information available on the Internet, to the ease of use of search engines to find information, and to the electronic submission (*cyberplagiarism*). There are even Internet platforms that prepare and offer free homework to be downloaded.

This article proposes several strategies to implement a program for personal use to detect one of the currently most widespread types of plagiarism as is copying and pasting fragments of texts from the Internet. These proposals are targeted to detect plagiarism in papers of different type, including reports, slides and web pages. This system returns an matching index for each delivery, thus teachers can clearly identify copies and focus efforts on reviewing only original works.

Palabras clave

Detección de plagio, memorias, diapositivas, páginas web.

1. Introducción

El incremento del plagio está en relación directa con el aumento de las exigencias académicas, el grado de procrastinación del estudiante, la abundancia de información disponible y su facilidad de reutilización, entre otras causas [9]. Sin embargo, la relación completa es mucho más exhaustiva y compleja [2] y en el conocimiento de la misma se encuentra parte de la solución del problema. Plagio ha existido siempre, pero nunca ha sido tan fácil recurrir a él.

Reutilizar trabajos y copiar y pegar textos sin ninguna transformación forma parte del pasado, estas prácticas han sido sustituidas por otras más sofisticadas. Actualmente existen diferentes tipos de plagio como los indicados en el trabajo de Tripathi[11]: plagiar textos sólo traduciéndolos, reutilizar ideas sin citar, o el último escalón, comprar trabajos con certificados de garantía de no ser detectados por programas antiplagio son algunas de las amenazas a las que han de enfrentarse los docentes que han de evaluar los trabajos realizados por sus alumnos.

La mayoría de los tipos de plagio son difíciles de detectar incluso para expertos debido a su sofisticación para enmascarar las ideas o párrafos originales de los que provienen. El problema se agrava cuando sobre él inciden variables que hay que tener muy en cuenta como por ejemplo la diferente formación tecnológica entre alumnos nacidos en plena era digital y profesores

que han permanecido en el conservadurismo tecnológico; gravedad que se incrementa cuando un porcentaje alto de las nuevas generaciones que acceden a la universidad vienen con sus prácticas de plagio realizadas y superadas en niveles educativos previos [10].

Mientras en el mundo anglosajón la prevención del plagio ha sido una medida aplicada desde hace años para tratar de garantizar el reconocimiento justo del trabajo realizado, en otros países, entre los que se encuentra España, esto no ha sucedido [7]. A la ausencia de medidas preventivas contra el plagio se suma un cierto retraso en la aplicación de elementos disuasorios, como la utilización de software de detección.

La prevención e identificación del plagio se encuentra en un estado incipiente en España. Incipiente porque el concepto de plagio todavía no está claro para los que han de evitarlo, porque el alcance del problema no ha sido suficientemente evaluado [5], porque el daño que esta práctica genera no es suficientemente valorado, y porque una vez detectado no existe una normativa clara para aplicar medidas correctoras.

Generalmente los autores coinciden en afirmar que la mejor medida para evitar el plagio es la concienciación para no cometerlo, pero hasta que este objetivo no se consiga la medidas disuasorias son esenciales, y en esto también estamos en un estado incipiente, pues no hay desarrollos informáticos nacionales que puedan competir con otros ya implantados en el mercado, ni directrices que ayuden a la adquisición de uno u otro producto comercial.

Los grandes desarrollos informáticos tienen una eficacia demostrada, pero presentan problemas que hacen dudar a los potenciales compradores: el precio, las condiciones de adquisición, y la obligación de depositar los contenidos a analizar en las bases de datos de los programas antiplagio. Además necesitan de un elevado presupuesto anual dedicado para usarlas habitualmente por todo el profesorado de un centro o universidad, o bien, se pueden implementar aplicaciones propias basadas en lenguaje natural como recomiendan en [6], pero estas necesitan tratamientos complejos de documentos, grandes corpus de datos para que sean eficaces, y además tienen un número de palabras limitadas para extraer características y detectar similitudes basadas en índices semánticos de latencia [4, 3].

También existen herramientas gratuitas para abordar el plagio pero presentan limitaciones que desaconsejan su utilización por diferentes motivos. Pese al reclamo de su gratuidad, esta no es tal. Tanto en sus versiones sin registro, como una vez registrado el usuario, el uso se limita a un tamaño definido de fichero o a un número de caracteres (*Plagiarisma, See sources, Plagiarism checker, DuplyChecker*). Otros ofrecen una revisión sin límite de extensión restringida a un número de artículos al día (*Dustball*). Sólo unos pocos analizan Wikipedia, y cuando lo hacen es en su versión inglesa (*Crot*). Otros productos se instalan en local pero no son multiplataforma (*Viper, Wcopyfind*), no tienen base de datos propia, no permiten el análisis simultáneo de varios documentos; son productos dirigidos al análisis de páginas web (con límite de resultados como el caso de *Copyscape*), imágenes o texto, pero no son una herramienta única para la detección de coincidencias en estos tres tipos de contenidos. En resumen, muchos de ellos son versiones demo del programa comercial.

Erradicar los distintos tipos de plagio académicos (se puede consultar la taxonomía en [11]) de forma automática no es posible, pero detectar un tipo de plagio literal (*Photocopy*) obtenido directamente de alguna fuente de Internet sí lo es, además este tipo de plagio es el más habitual.

Las estrategias que presentamos en este artículo intentan superar estas limitaciones: la idea no es implementar un producto que sirva de reclamo a una versión comercial, ya que ésta no existiría; permitiría la revisión sin límite de caracteres o tamaño del fichero; analizaría no solo Wikipedia, sino también otras bases de datos especializadas como ResearchGate, Academia o los repositorios institucionales abiertos. En nuestro caso, el prototipo se ha implementado usando el lenguaje de programación JAVA dado que es multiplataforma, ha sido probado en Windows, Linux y Mac, ofrece al usuario la opción de crear su propia base de datos, permite el análisis simultáneo de varios documentos, y por último puede analizar textos, páginas web y presentaciones sin los límites descritos anteriormente para otros productos. Remarcar además que la intención principal de este artículo es presentar las estrategias básicas que se han de seguir para realizar un desarrollo que se adecue a las necesidades de cada uno.

En la siguiente sección se describe en detalle el problema que se pretende resolver y las soluciones propuestas. Estos algoritmos han sido aplicados para la detección de plagio en los trabajos realizados para una asignatura durante tres años. Los resultados extraídos de la experimentación se analizarán en el apartado 3 y por último se exponen las conclusiones obtenidas.

2. Propuesta de estrategias para la detección de plagio

En esta sección se presentan las dos estrategias propuestas para la detección de plagio. Como el tipo de los trabajos sobre los que se puede aplicar puede ser muy variado, se van a diferenciar entre los dos siguientes casos:

- *Trabajos de texto de tipo memoria, resumen o informe.* En este tipo de documentos podemos realizar búsquedas literales de fragmentos suficien-

temente extensos para detectar una coincidencia exacta. Por lo tanto se podrá detectar el plagio literal de fragmentos extensos de texto copiados de alguna fuente accesible, normalmente Internet (Wikipedia, foros, blogs, etc.), o bien, algún trabajo previo ya entregado.

- *Presentaciones en formato diapositivas o páginas web.* En esta aproximación la detección de plagio es diferente a la anterior ya que habitualmente una diapositiva contiene un número reducido de palabras y si intentamos una búsqueda literal obtendremos numerosos resultados que no se pueden considerar plagio. La solución es buscar diapositivas o páginas web entre otros trabajos previos o del resto de compañeros que se parezcan al contenido actual y devolver un índice de similitud con los trabajos anteriores.

En ambos casos deberemos agrupar los documentos entregados por los alumnos en uno o varios directorios, de forma que con un sola llamada o ejecución se realicen todas las comprobaciones y se obtenga un informe detallado con los resultados. Este informe contendrá un índice de concordancia (entre 0 y 1) para indicar cuán parecidos son los contenidos examinados con respecto al resto de entregas, o bien, a otros existentes en Internet.

En las siguientes secciones se detallan los algoritmos seguidos para la detección de similitudes para cada uno de los casos.

2.1. Plagio literal de fragmentos de texto

Esta aproximación sería adecuada para trabajos tipo memoria, resumen o informe, de los que podemos extraer párrafos suficientemente extensos como para compararlos de forma literal. La idea básica es aprovechar la potencia que nos ofrecen los buscadores de Internet habituales (como *Bing*, *Google*, *Yahoo*, *DuckDuckGo*, *IxQuick*, etc.) para encontrar fragmentos literales de texto en páginas web de Internet.

La solución propuesta es:

1. Convertir el trabajo a examinar a formato texto. Desde la misma herramienta que se creó, una compatible o bien, si el trabajo está en formato PDF convertirlo a texto con la herramienta libre *pdftotext* (`http://www.foolabs.com/xpdf/`);
2. Leer el texto por párrafos y dividirlos en fragmentos consecutivos de palabras (mínimo de 70 caracteres por fragmento, ya que es un tamaño suficiente para evitar los falsos negativos en la búsquedas, y además, lo soportan los principales buscadores);
3. Buscar cada fragmento en el mismo orden (búsqueda literal) en un buscador de Internet,
   - si devuelve resultados, el fragmento pertenece a otros textos en Internet;
   - sino el fragmento es considerado original;
4. Devolver un índice de coincidencias final (nº de fragmentos encontrados / nº total de fragmentos buscados) junto a un informe con los fragmentos encontrados en Internet y, al menos, una referencia al primer sitio web que lo contiene. Estas indicaciones sirven como base al profesor para determinar si partes del texto son consideradas plagio o no.

Esta propuesta tiene el inconveniente de que los buscadores detectan que se realizan muchas peticiones desde la misma dirección IP, o el mismo navegador y restringen el número de accesos. Una solución sencilla es insertar una pequeña espera aleatoria en el programa que realiza la llamada, o bien, contratar un límite de búsquedas. En la actualidad el buscador *Bing de Microsoft* (`https://www.bing.com/`) es el recomendado para realizar las búsquedas, ya que es el que menos restricciones aplica al número de búsquedas abiertas[1] y los resultados literales son de buena calidad.

2.2. Plagio por similitud entre conjuntos de palabras

En este apartado se consideran las entregas tipo presentación de diapositivas o páginas Web. En ellas abundan los párrafos cortos y si se utilizara la aproximación del apartado anterior se obtendrían considerables falsos positivos ya que las longitudes de sus fragmentos serían excesivamente cortas.

Para solucionar este problema se propone una aproximación distinta a la anterior en la que se aprovecha el concepto de página o diapositiva y se utiliza toda la información textual incluida en la misma para realizar las búsquedas. De esta forma se puede realizar la comparación utilizando técnicas de similitud o distancia para detectar otras páginas o diapositivas similares.

Como medida de similitud entre dos diapositivas o páginas Web se ha utilizado el índice de Jaccard (J) [8] para la comparación:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

Donde A y B son los conjuntos de palabras extraídos de cada una de las diapositivas o páginas Web a comparar.

[1] También dispone de una versión por suscripción llamada *Bing Search API* que tiene tramos escalonados de precios según el límite de búsquedas mensuales. Cabe destacar que el primer tramo es gratuito y permite realizar hasta 5000 búsquedas que podrían ser suficientes para un uso esporádico.

Para obtener el índice de coincidencia de una entrega se tendrían que realizar los siguientes pasos:

1. Recorrer todas las diapositivas o páginas del trabajo actual y calcular para cada una el mínimo índice de Jaccard con respecto a otra presentación o página web almacenada.
2. Promediar los índices de Jaccard calculados anteriormente para obtener la similitud entre la entrega actual y otra almacenada.
3. Repetir los pasos 1 y 2 para cada entrega almacenada, y finalmente obtener la entrega que más se parezca a la actual con su índice de similitud correspondiente.

En el algoritmo implementado se ha utilizado además un código Hash [1] para guardar la representación de las palabras. De esta forma se consigue que el tamaño de las presentaciones guardadas (vectores de vectores de enteros que representan diapositivas con sus palabras) ocupen menos y las comprobaciones sean más rápidas.

3. Caso de estudio

Los experimentos se han realizado en la asignatura de DESARROLLO CURRICULAR Y AULAS DIGITALES EN EDUCACIÓN PRIMARIA (DCADEP) perteneciente al GRADO EN MAESTRO EN EDUCACIÓN PRIMARIA de la Universidad de Alicante. En esta asignatura partimos del siguiente contexto:

- Los profesores en general no usan programas antiplagio.
- Los alumnos usan la técnica de cortar y pegar párrafos (sin citas) en numerosos trabajos sin que ello suponga depreciación en su nota o apercibimiento por parte del profesor.

Los trabajos usados para la experimentación pertenecen a los cursos académicos 2013-2014 (2013), 2014-2015 (2014) y 2015-2016 (2015) e incluyen memorias realizadas sobre temas determinados (fragmentos de texto extensos), presentaciones (diapositivas, fragmentos de texto cortos) y diseño de páginas web (fragmentos de texto cortos).

Es importante destacar que en el curso 2014 y 2015 se avisó desde el principio que las memorias debían tener una redacción original y que se usaría un programa antiplagio para detectar casos de copia. En el caso del curso 2013 no se avisó y tampoco se usó ningún programa antiplagio. Los datos sobre el índice de coincidencia entre trabajos se han calculado posteriormente para su comparación y utilización en este artículo.

El cuadro 1 resume las tasas de coincidencia en los trabajos entregados junto a algunas características adicionales como la nota obtenida en el trabajo, el número de componentes del mismo, curso académico y turno.

Como vemos en los valores estadísticos del Cuadro 2 y gráficamente en la Figura 1, la media y la mediana de coincidencias en el curso 2013 son superiores a las de los cursos 2014 y 2015, y también su rango intercuartílico (diferencia entre sus cuartiles tercero y primero) lo que significa que las memorias contienen menos fragmentos coincidentes en el 2014 y 2015. Es interesante observar como el hecho de haber avisado a los alumnos produce una reducción en el índice de coincidencia medio pase de un 10 % a un 6.5 % y 3.7 % en el caso del turno de mañana y de un 13.1 % a un 3.6 % y 6.9 % en el turno de tarde.

En la Figura 2 se detectan claramente 4 casos atípicos[2] en el curso 2013, 3 en el turno de mañana (17 alumnos) y 1 en el de tarde (2 alumnos);en el curso 2014 hubo 3 casos atípicos 2 de mañana (8 alumnos) y 1 de tarde (5 alumnos) y en el curso 2015 ninguno en el turno de mañana y 2 (9 alumnos) en el turno de tarde.

Respecto al rango intercuartílico en el índice de coincidencia en los cursos que se usó el programa antiplagio, 2014 y 2015, se observa que en el turno de tarde en el 2015 es superior al resto. Claramente al incrementarse el número de alumnos que hicieron caso omiso de la advertencia de realizar una redacción propia y original en las memorias.

La Figura 3 representa gráficamente la relación entre el índice de coincidencia y la nota obtenida en el trabajo. Esta relación está representada mediante una regresión lineal simple y su confianza en cada punto de la misma. La situación ideal sería un recta descendente, es decir a mayor coincidencia con otros trabajos, menor nota, o bien, que fuera despreciable. Como se puede ver en la figura, en el curso 2013 la recta de regresión se mantiene estable en ambos turnos, lo que significa que la nota es independiente del índice. Sin embargo, en el turno de mañana del curso 2014 la línea es ligeramente ascendente. Este comportamiento es debido a que se avisó a varios grupos sobre su alto índice de coincidencia[3], por lo que revisaron sus entregas y finalmente obtuvieron mejores notas. En el turno de tarde de 2014, a pesar de haberles avisado, los trabajos no fueron modificados, o bien, su actualización fue mínima, por lo que claramente la recta de regresión es descendente. El caso ideal se encuentra en el turno de mañana del curso 2015, donde el índice de coincidencia está por debajo del 10 % y además la recta de regresión es descendente. En el caso del turno de tarde los índices de coincidencia son mayores pero se mantiene la tendencia inversa entre la nota obtenida y el

[2] Siguiendo la definición general, se ha considerado como caso atípico a los valores que exceden $Q3 + (Q3 - Q1) * 1{,}5$ siendo Q los cuartiles de la distribución.

[3] Debido a que se trataba de un sistema en fase de pruebas decidimos adoptar la estrategia de avisarles.

			Mañana			Tarde		
	Grupo	**Unidad**	**Nota**	**Índ. coincidencia**	**Componentes**	**Nota**	**Índ. coincidencia**	**Componentes**
2013	**A**	01	7,4	17 % (27/157)	6	9,0	11 % (61/536)	3
		02	8,7	4 % (6/137)	4	8,3	9 % (17/180)	3
		03	8,9	10 % (17/174)	6	8,3	14 % (16/118)	2
		04	7,4	8 % (13/158)	5	7,5	17 % (26/152)	4
		05	9,9	10 % (14/145)	4	8,6	**36** % (76/209)	2
		06	8,3	3 % (4/120)	5	9,3	9 % (19/207)	3
		07	8,1	2 % (2/90)	5	9,5	5 % (17/367)	3
		08	7,8	3 % (4/115)	6	8,0	11 % (22/202)	3
		09	8,1	7 % (3/42)	4	8,5	10 % (23/225)	4
		10	9,1	**31** % (61/196)	6	8,2	28 % (50/180)	1
		11	8,5	11 % (18/169)	4	9,4	3 % (5/170)	3
	B	01	9,8	5 % (6/119)	5	9,6	19 % (26/139)	4
		02	9,2	6 % (6/96)	4	7,9	5 % (6/126)	4
		03	9,5	1 % (1/71)	5	8,3	7 % (10/152)	4
		04	8,4	3 % (3/113)	4	6,5	14 % (22/155)	3
		05	7,4	**26** % (31/117)	6	9,1	2 % (2/120)	4
		06	9,9	10 % (21/207)	4	8,8	14 % (15/110)	4
		07	8,9	0 % (0/84)	4	9,0	26 % (57/217)	5
		08	8,6	13 % (20/157)	6	5,8	5 % (4/82)	4
		09	9,0	**34** % (42/123)	5	7,2	1 % (1/86)	6
		10	7,0	12 % (22/183)	4	8,0	29 % (53/180)	5
		11	6,8	3 % (3/107)	3	8,0	13 % (15/120)	3
2014	**A**	01	10,0	4 % (4/100)	5	6,8	2 % (2/87)	5
		02	9,9	5 % (5/100)	5	4,5	6 % (6/100)	4
		03	9,1	3 % (3/79)	5	7,3	4 % (7/167)	3
		04	10,0	9 % (8/91)	6	8,3	2 % (2/97)	3
		05	9,8	4 % (4/100)	6	8,3	2 % (2/97)	6
		06	8,1	7 % (11/162)	3	5,5	4 % (5/118)	6
		07	9,8	6 % (4/61)	5	7,5	0 % (0/68)	3
		08	7,8	1 % (1/82)	3	8,5	5 % (5/91)	4
		09	8,5	1 % (1/73)	5	6,0	2 % (2/111)	5
		10	8,1	**18** % (24/133)	4	5,4	3 % (3/89)	2
		11	8,3	5 % (6/120)	6	8,0	0 % (0/52)	5
	B	01	9,6	9 % (10/110)	6	9,8	5 % (5/106)	4
		02	7,5	4 % (5/109)	5	3,8	2 % (1/62)	4
		03	9,0	1 % (1/73)	3	7,5	6 % (4/63)	4
		04	6,8	2 % (3/125)	6	8,8	0 % (0/100)	6
		05	7,6	7 % (5/67)	5	8,8	0 % (0/33)	6
		06	8,1	**24** % (19/79)	4	9,5	6 % (6/94)	5
		07	9,5	1 % (1/104)	6	4,6	2 % (2/85)	5
		08	8,6	3 % (2/68)	5	7,8	0 % (0/51)	4
		09	7,3	5 % (4/72)	5	8,0	3 % (2/68)	5
		10	8,3	13 % (20/156)	5	6,8	3 % (3/87)	6
		11	8,5	12 % (13/108)	3	5,0	**22** % (23/104)	5
2015	**A**	1	8,8	4 % (5/131)	5	4,5	21 % (26/122)	5
		2	10,0	1 % (2/149)	5	8,4	1 % (1/95)	4
		3	8,6	1 % (2/195)	5	7,9	2 % (2/104)	3
		4	8,1	8 % (11/146)	6	6,6	16 % (21/134)	3
		5	8,4	6 % (7/112)	6	10,0	1 % (1/133)	6
		6	7,2	1 % (2/142)	3	7,9	1 % (1/118)	6
		7	6,9	4 % (4/103)	5	5,5	0 % (0/43)	3
		8	8,3	1 % (1/119)	3	7,8	6 % (5/79)	4
		9	8,9	0 % (1/222)	5	7,8	6 % (5/90)	5
		10	7,8	8 % (11/136)	4	6,4	10 % (11/113)	2
		11	9,9	3 % (5/156)	6	7,0	**25** % (15/60)	5
	B	1	8,6	5 % (7/144)	6	8,5	8 % (10/120)	4
		2	9,7	1 % (2/159)	5	10,0	2 % (4/183)	4
		3	9,5	0 % (0/124)	3	5,7	1 % (1/86)	4
		4	9,8	3 % (4/134)	6	6,4	11 % (16/147)	6
		5	8,2	4 % (7/157)	5	6,9	2 % (2/100)	6
		6	8,4	2 % (3/131)	4	5,5	8 % (9/120)	5
		7	6,5	7 % (7/103)	6	7,1	3 % (3/90)	5
		8	8,6	2 % (2/91)	5	7,7	4 % (2/53)	4
		9	9,6	2 % (2/96)	5	8,2	0 % (0/157)	5
		10	9,0	8 % (6/80)	5	7,6	**22** % (42/194)	6
		11	7,1	10 % (16/160)	3	8,4	2 % (3/148)	5

Cuadro 1: Información sobre las memorias entregadas durante los cursos 2013, 2014 y 2015 sobre un total de 577 alumnos, 182, 206 y 189, respectivamente. Los trabajos con un índice de coincidencia atípico están marcados en negrita (en la figura 2 se pueden ver dichos casos atípicos de forma gráfica).

		Mañana		Tarde	
		Coincidencia	**Nota**	**Coincidencia**	**Nota**
2013	**Mínimo**	0,0 %	6,8	1,0 %	5,8
	Cuartil 1	3,0 %	8,2	5,5 %	8,0
	Mediana	7,5 %	8,6	11,0 %	8,4
	Media	**10,0** %	**8,6**	**13,1** %	**8,5**
	Cuartil 3	11,8 %	9,3	16,3 %	9,5
	Máximo	34,0 %	9,9	36,0 %	10,0
2014	**Mínimo**	1,0 %	4,5	0,0 %	3,8
	Cuartil 1	3,0 %	6,9	2,0 %	5,5
	Mediana	5,0 %	7,8	2,5 %	7,8
	Media	**6,5** %	**7,6**	**3,6** %	**7,2**
	Cuartil 3	8,5 %	8,5	4,8 %	8,8
	Máximo	24,0 %	9,8	22,0 %	9,8
2015	**Mínimo**	0,0 %	6,5	0,0 %	4,5
	Cuartil 1	1,0 %	8,1	1,2 %	6,4
	Mediana	3,0 %	8,6	3,5 %	7,7
	Media	**3,7** %	**8,5**	**6,9** %	**7,4**
	Cuartil 3	5,7 %	9,3	9,5 %	8,1
	Máximo	10,0 %	10,0	25,0 %	10,0

Cuadro 2: Valores estadísticos sobre el índice de coincidencia (de 0 a 100 %) y la nota (de 0 a 10). Se destaca en negrita el valor de la media por curso.

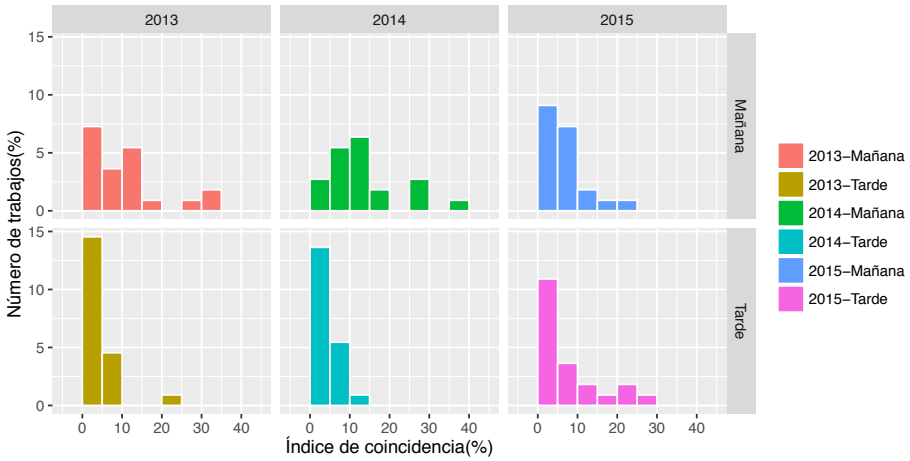

Figura 1: Histogramas sobre el índice de coincidencia según turno y curso académico.

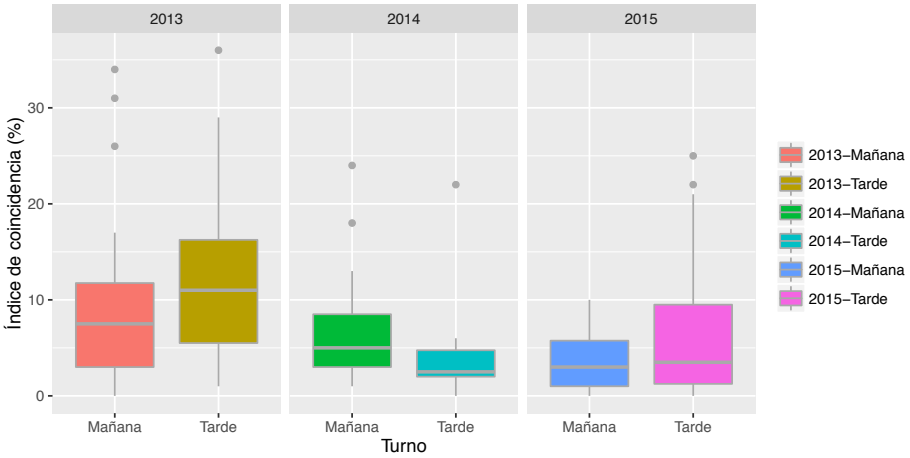

Figura 2: Gráfico de cajas de los índices de coincidencia según grupos y curso académico.

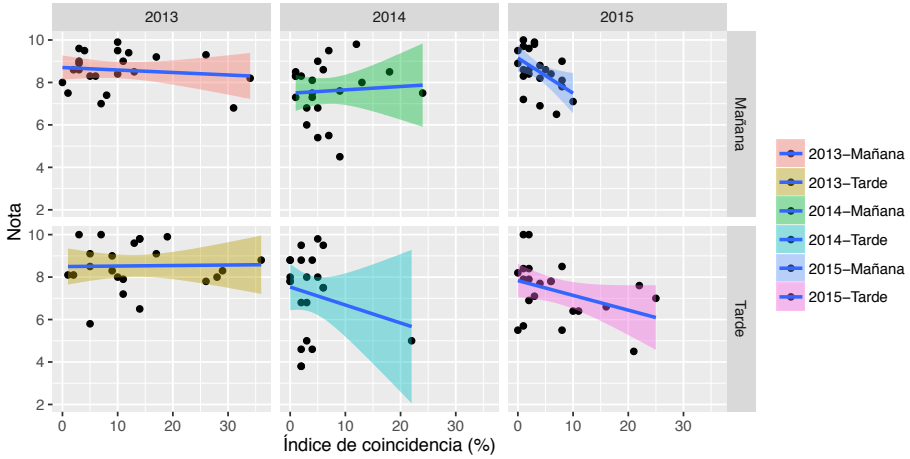

Figura 3: Gráfico que relaciona con una regresión lineal simple la nota obtenida en un trabajo y el índice de coincidencia del mismo para los diferentes grupos y cursos.

índice de coincidencia.

Por el contrario, no se detectaron índices significativos de coincidencia en diapositivas y/o páginas web en esta asignatura. Seguramente debido al planteamiento de las mismas, ya que en el caso de las diapositivas primero tenían que realizar el trabajo desarrollado en formato texto (memoria) para realizar la presentación en diapositivas. En el caso de la elaboración de páginas web los trabajos estaban relacionados con la confección de Webquest muy concretas. Además, en esta asignatura impartimos docencia cinco profesores en diferentes turnos. Esta circunstancia parece influir positivamente en el número de plagios, ya que hace que la copia sea más difícil debido a la diversidad de los trabajos requeridos por los distintos profesores.

4. Conclusiones y trabajos futuros

En este trabajo se han presentado dos aproximaciones sencillas basadas en texto para calcular un índice de coincidencias en el caso de búsquedas literales y otro para similitud de términos.

Este tipo de estrategias para la detección de plagios en actividades basadas en texto son imprescindibles actualmente. Suponen un ahorro económico importante para las universidades, ya que no se pagan licencias de uso de programa antiplagio, evitan subir a Internet los trabajos completos de los alumnos para que sistemas gratuitos o de pago lo revisen, son fáciles de programar en el lenguaje que prefiramos, y además permiten que el profesor se centre únicamente en la corrección de trabajos originales. Este sistema también sirve para recabar información sobre qué fragmentos no son originales y de qué páginas web o trabajos se han extraído para mostrarlos en caso necesario.

Como se deduce de nuestro caso de estudio, el hecho de advertir y usar un programa antiplagio reduce el promedio del índice de coincidencia en los textos a evaluar y claramente se detectan los casos más importantes (atípicos) que hacen caso omiso a las recomendaciones para actuar en consecuencia. Por ello las dos estrategias propuestas supondrían una reducción drástica de este tipo de prácticas entre el alumnado, con un coste mínimo para las universidades que las aplicaran y facilidad para su utilización por parte del profesorado.

Como trabajos futuros se podrían extender las estrategias a trabajos basados en imágenes, así como aplicar la idea de buscar los fragmentos en varios buscadores para balancear la carga de búsquedas, y además, obtener resultados más diversos, ya que algunos de ellos no encuentran fragmentos que otros sí lo hacen, a la vez que se evitarían los límites de búsquedas impuestos por un solo buscador. Otra opción interesante sería ampliar las búsquedas a bases de datos privadas contratadas por los centros educativos o universidades.

Referencias

[1] Burton H Bloom. Space/time trade-offs in hash coding with allowable errors. *Communications of the ACM*, 13(7):422–426, 1970.

[2] Nina C. Heckler and David R. Forde. The role of cultural values in plagiarism in higher education. *Journal of Academic Ethics*, 13(1):61–75, 2015.

[3] R Kharat. Semantically detecting plagiarism for research papers [online]. *International Journal of Engineering Research and Applications (IJERA)–2013*, (3):3.

[4] Todd A Letsche and Michael W Berry. Large-scale information retrieval with latent semantic indexing. *Information sciences*, 100(1):105–137, 1997.

[5] A. Cayuela Mateo, A. Tauste Francés, M.M. Seguí Crespo, J.M. Esteve Faubel, and E. Ronda Pérez. ¿cómo medir el plagio entre alumnos universitarios? revisión de instrumentos utilizados en artículos científicos. In *XIII Jornadas de Redes de Investigación en Docencia Universitaria: Nuevas estrategias organizativas y metodológicas en la formación universitaria para responder a la necesidad de adaptación y cambio*, pages 210–216. Universidad de Alicante, 2015.

[6] Maxim Mozgovoy, Tuomo Kakkonen, and Georgina Cosma. Automatic student plagiarism detection: future perspectives. *Journal of Educational Computing Research*, 43(4):511–531, 2010.

[7] Enrique Ortega Martínez. La respuesta al plagio en la educación superior: un estudio internacional. Publicación en CD:15, 2011.

[8] Raimundo Real and Juan M Vargas. The probabilistic basis of jaccard's index of similarity. *Systematic biology*, 45(3):380–385, 1996.

[9] Jaume Sureda-Negre, Rubén Comas-Forgas, and Mercè Morey. Las causas del plagio académico entre el alumnado universitario según el profesorado. *Revista Iberoamericana de Educación*, (50):197–220, 2009.

[10] Jaume Sureda-Negre, Rubén Comas-Forgas, and Miquel Francesc Oliver-Trobat. Plagio académico entre alumnado de secundaria y bachillerato: Diferencias en cuanto al género y la procrastinación. *Comunicar: Revista Científica de Comunicación y Educación*, 22(44):103–111, 2015.

[11] Richa Tripathi and S. Kumar. Plagiarism: A plague. In *7th International CALIBER-2009*. INFLIBNET Center, 2009.