Revista Cubana de Ciencias Informáticas
ISSN: 2227-1899 | RNPS: 2301
Vol. 8, No. 4, Octubre-Diciembre, 2014
Págs. 130-144
http://rcci.uci.cu

Tipo de artículo: Artículo original
Temática: Inteligencia artificial
Recibido: 28/03/2014 | Aceptado: 30/09/2014

Técnica de clasificación bayesiana para identificar posible plagio en información textual

Bayesian classification technique to identify possible plagiarism in textual information

Grethell Castillo Reyes [1*], Yanisley González González [2], Guillermo Luzua Farias[1]

[1] Centro de Desarrollo Geoinformática y Señales Digitales (GEYSED). Universidad de las Ciencias Informáticas, Carretera a San Antonio de los Baños, km 2 ½, Torrens, Boyeros, La Habana, Cuba. CP.: 19370.
[2] Centro de Información Científico Técnico (CICT). Universidad de las Ciencias Informáticas, Carretera a San Antonio de los Baños, km 2 ½, Torrens, Boyeros, La Habana, Cuba. CP.: 19370.

[*] Autor para correspondencia: gcreyes@uci.cu

Resumen

En la rama de la educación, el plagio de documentos es un problema con tendencia a ir en aumento. En el ámbito de la investigación científica, la manifestación de trabajos investigativos plagiados ha estado extendiéndose, por lo que se ha hecho necesaria la búsqueda de soluciones para contrarrestar este problema. En particular la Universidad de las Ciencias Informáticas decidió implementar un sistema informático que permita verificar la existencia de plagio en los artículos científico – técnicos a publicar en los sistemas de información académica que allí se manejan: Serie Científica, Revista Cubana de Ciencias Informáticas y Repositorio Institucional fundamentalmente. El objetivo de este trabajo es proponer el uso de varias técnicas de detección de plagio para el sistema a desarrollar, así como la utilización de un método de aprendizaje automático para la clasificación de los documentos sospechosos.

Palabras clave: aprendizaje automático, clasificación, detección, documentos, plagio.

Abstract

In the field of education, the document plagiarism is a problem with tendency to go up. In the field of scientific research, has been demonstrated that the plagiarized research papers has been spreading, so it has become necessary to find a

solution to counteract this problem. In particular the University of Informatics Sciences decided to implement a computerized system to check for plagiarism in scientific articles for publishing in the academic information systems of the university such as: Scientific Series, Cuban Journal of Computer Science and the Institutional Repository fundamentally. The aim of this paper is to propose the use of various techniques for the plagiarism detection system to be developed, and the use of a machine learning method for the classification of suspicious documents.
Keywords: *classification, detection, documents, machine learning, plagiarism.*

Introducción

El plagio es una mala práctica que ha existido siempre, pero no en gran medida. Tener acceso a grandes volúmenes de información en la actualidad, para muchos, se ha convertido en un beneficio a la hora de buscar trabajos ya realizados por otros para tratar de imitarlos. En estos tiempos, este fenómeno se manifiesta con mayor fuerza debido a que con la presencia de Internet, se hace más variada, directa y accesible la búsqueda de documentación.

La (IEEE, 2014) define plagio como la "reutilización de las ideas, procesos, resultados o palabras de alguien más sin mencionar explícitamente a la fuente y su autor". Para nada es extraño encontrar documentos que no han sido escritos originalmente (parcial o totalmente) por quienes dicen serlo. Cuando se copia la idea de alguien sin hacer referencia a su autoría se está cometiendo el delito de plagiar. Por su parte, el plagio de documentos, según (Cedeño, 2008), se define como el plagio que implica incluir fragmentos de texto que se encuentran en documentos escritos por otro autor en un documento propio sin incluir el crédito correspondiente.

Al presentar una obra ajena haciéndola pasar como propia no solo se incurre en delitos penalizados por la ley, sino que también se ve afectada la ética moral y profesional del individuo que incurre en este delito. En el ámbito de la investigación científica, este problema también va en aumento. En su investigación, (Vega, 2011) afirma que la manifestación de trabajos investigativos plagiados ha estado extendiéndose en los comités de publicación de revistas de alto prestigio. Es por ello que se ha hecho necesaria la búsqueda de una solución para contrarrestar este problema.
A partir de esto es que surge la importancia absoluta de utilizar herramientas informáticas que permitan la detección automática de plagio. No siempre lo plagiado de un texto es lo copiado exactamente, muchas veces se cambian palabras por sinónimos, se intercambian ideas de una parte de la frase a otra, entre distintas transformaciones que complican la detección del plagio. Aunque los profesionales incurran en este fraude por ignorancia, no deja de repercutirle en su formación, lo que puede traerle consigo la pérdida de credibilidad y prestigio profesional.

En la Universidad de las Ciencias Informáticas (UCI) existen varios sistemas de información académica que permiten la publicación de investigaciones científicas, entre ellos: el Repositorio Institucional, la Serie Científica de la Universidad (SC – UCI) y la Revista Cubana de Ciencias Informáticas (RCCI). Así mismo cuenta con los sitios de gestión de eventos como pueden ser: UCIENCIA, Fórum de Ciencia y Técnica, Jornadas Científicas Estudiantiles, Peñas Tecnológicas, entre otros. Antes de aceptar las investigaciones enviadas por los autores para publicar, estas son remitidas a un colectivo de revisores y sometidas a un proceso de arbitraje en el que se verifican una serie de requisitos que se deben cumplir.

Hoy, este proceso de revisión y arbitraje no contempla oficialmente la verificación de coincidencias en el contenido de los artículos enviados para revisión, con el contenido de otros ya publicados. En algunos casos, por gestión propia de los revisores y mediante la utilización de su cuota de navegación en Internet, se utilizan herramientas en línea que permiten analizar hasta qué punto el material enviado coincide con otra colaboración publicada o en proceso de publicación, con el objetivo de analizar si existe posible plagio.

Lo anterior trae como consecuencia que:

(1) Con el uso de herramientas en línea no se detectan las coincidencias con documentos que no han sido publicados en internet.
(2) A través del uso de herramientas en línea se envíe información sensible a programas de terceros comprometiendo la confidencialidad de la información.
(3) Se consuma ancho de banda por la utilización de herramientas en línea.

A partir de las problemáticas planteadas anteriormente y del estudio realizado acerca de los sistemas que permiten la detección de plagio en documentos digitales de texto, la Universidad decidió implementar una herramienta local que permita analizar la coincidencia de un documento en revisión con otros ya publicados en sistemas de información académica, con el objetivo de detectar posible plagio en el mismo. Por lo general, las herramientas de este tipo a nivel global utilizan algoritmos a partir de los cuales se deduce si un documento determinado contiene pasajes sospechosos de un texto original. En la mayoría de los casos, estos algoritmos aplicados de manera individual tienen sus limitantes, ya que unos contemplan conceptos o rasgos de la detección de plagio que otros no. Por tal motivo, en este trabajo se propone la utilización de varios algoritmos de detección de plagio para el sistema a desarrollar. La propuesta tiene el

objetivo de valorar al unísono el comportamiento de varios rasgos en los documentos sospechosos de plagio. Para lograrlo, se plantea la utilización de un método de aprendizaje automático: el algoritmo Naïve Bayes, el cual es aplicado para la clasificación de los documentos según los resultados emitidos por los algoritmos.

Metodología computacional

La tarea de la detección de plagio según (Cedeño, 2008) consiste en que, dado un conjunto *V* de documentos originales y un documento sospechoso *s*, realizar una comparación entre *s* y el conjunto *V* para determinar si *s* contiene fragmentos plagiados de algún$v_i \in V$. El proceso que se propone en el siguiente trabajo contempla dos fases fundamentales para concluir si un documento es plagiado o no. La primera de ellas es la aplicación de varias técnicas de detección de plagio al documento que se desea verificar: el análisis basado en n-gramas, el análisis basado en el modelo de espacio vectorial y el cálculo de la máxima subsecuencia común. La segunda fase, consiste en emplear los resultados emitidos en la fase anterior para clasificar el documento en "Plagiado" o "No plagiado", utilizando un algoritmo de aprendizaje automático.

Rasgos observados en la detección de plagio en texto

Autores como (Clough, 2003) y (Vega, 2011) definen algunas características o rasgos comunes presentados por las obras plagiadas. A continuación se mencionan y se describen brevemente algunas de ellas que servirán como base de la investigación. Las mismas han motivado la implementación de algoritmos y métodos de detección de plagio que se describen posteriormente.

(1) Distribución de palabras. La distribución de las palabras se refiere a su frecuencia o habitualidad en determinado documento. Cada autor prefiere el uso de ciertos términos en lugar de otros; por lo que encontrar varias palabras que sean usadas con la misma frecuencia da pie a pensar que dichos textos están influenciados y que podría tratarse de un plagio.

(2) Secuencias de texto común. Los textos escritos de forma independiente no deberían contener secuencias (de palabras o caracteres) comunes de gran longitud, incluso si abordan el mismo tema.

(3) Cantidad de texto común. Es habitual que documentos que abordan el mismo tema (inclusive aquellos que sólo son de temas relacionados) compartan cierta cantidad de texto, básicamente nombres y términos específicos del área. Pero si se trata de documentos escritos de forma independiente, esta cantidad de texto similar o idéntico debería ser pequeña.

Algoritmos y métodos para la detección de plagio en texto

A nivel global existen varios algoritmos y métodos que, bajo diferentes conceptos, permiten detectar las coincidencias existentes entre dos documentos. A continuación se detallan algunas de las características particulares de cada uno de los algoritmos y métodos que más resaltan en el estudio realizado, a partir de los rasgos enunciados anteriormente:

Análisis basado en n-gramas de palabras

Para realizar una estrategia de búsqueda flexible, (Pinto et al., 2011) basan la comparación de documentos en los n-gramas contenidos en ellos. Según (Vega, 2011), los n-gramas son trozos de *n* palabras del texto. El empleo de estos proviene de los modelos de lenguaje y su utilización en el reconocimiento del habla. Los métodos basados en n-gramas tienen la misma estructura: se toman n-gramas del documento en general de forma superpuesta, lo cual hace que la cantidad de n-gramas de un texto de *r* palabras sea igual a $r - n + 1$.

La correcta selección de los n-gramas de un documento es muy importante. Por ejemplo, si se eligen trozos demasiado pequeños, la probabilidad de que se repitan en otros textos será muy grande, sin importar que sean textos independientes (sin plagio). Por otro lado, elegir trozos muy grandes disminuye la posibilidad de que se encuentren en otro documento (Manchego, 2010) y las pequeñas modificaciones o reescrituras como la omisión o cambio de alguna palabra, evitaría que las porciones plagiadas fueran detectadas (Stamatatos, 2011). Este método suele ser combinado con otros métodos de análisis más detallados. En el caso de aplicar esta técnica se consideran los rasgos (2) y (3). El siguiente ejemplo muestra cómo se aplica la técnica basada en n-gramas de palabras, con *n* = 3. Los n-gramas que coinciden en ambos textos se muestran resaltados.

- Texto 1[1]: “Plagiar es copiar en lo sustancial obras ajenas, dándolas como propias”.
- Texto 2 (sospechoso): “Plagiar es reusar en lo sustancial palabras ajenas, dándolas como propias”.

N-gramas del Texto 1: [plagiar es copiar], [es copiar en], [copiar en lo], **[en lo sustancial]**, [lo sustancial obras], [sustancial obras ajenas], [obras ajenas dándolas], **[ajenas dándolas como]**, **[dándolas como propias]**.

N-gramas del Texto 2: [plagiar es reusar], [es reusar en], [reusar en lo], **[en lo sustancial]**, [lo sustancial palabras], [sustancial palabras ajenas], [palabras ajenas dándolas], **[ajenas dándolas como]**, **[dándolas como propias]**.

Modelo de espacio vectorial

[1]Tomado de (RAE, 2014)

Revista Cubana de Ciencias Informáticas
ISSN: 2227-1899 | RNPS: 2301
Vol. 8, No. 4, Octubre-Diciembre, 2014
Págs. 130-144
http://rcci.uci.cu

El modelo de espacio vectorial es otro de los métodos utilizados para la detección de plagio. Basa su funcionamiento en la representación del contenido de los documentos en términos de vectores. Posteriormente, mediante fórmulas matemáticas, arroja los resultados de las similitudes (Vega, 2011). Según este modelo, cada expresión del lenguaje natural puede representarse como un vector de pesos de términos, o la unidad mínima de información, como una palabra o la raíz sintáctica de una palabra. Para determinar la similitud que existe entre un documento y una consulta se calcula la distancia que existe entre los vectores que los representan (Zechner et al., 2009).

Un algoritmo basado en este método es el llamado *Word Chunking Overlap* (WCO por sus siglas en inglés), que calcula la similitud entre dos documentos utilizando la fórmula del coseno. Como resultado, se obtiene el valor del ángulo entre los vectores que representan los mismos. Mientras más pequeño sea el ángulo, más similares serán estos documentos. La desventaja de esta técnica radica en que el cambio del orden de las palabras puede cambiar el sentido de una oración, y precisamente este es un hecho que no se toma en cuenta. Por lo que para hacerlo efectivo es necesario combinarlo con alguna otra técnica, por ejemplo con el análisis de n-gramas de palabras. La aplicación de esta técnica cubre el rasgo (1) descrito en el epígrafe anterior. A continuación se muestra un ejemplo de cómo se aplica el método.

Suponiendo que se tienen los siguientes textos de referencia:

- Texto 1[2]: "Plagiar es reusar las ideas, procesos, resultados o palabras de alguien más sin mencionar explícitamente a la fuente y su autor".
- Texto 2: "Plagiar es copiar en lo sustancial obras ajenas, dándolas como propias".

Términos: [plagiar, reusar, copiar, inteligencia, ideas, artificial, procesos, alguien, palabras, fuente, autor, obras, propias, aprendizaje].

A partir de los términos anteriores se representan los vectores asociados a cada texto. Suponiendo que los pesos se asignen de la siguiente manera: 1 si aparece el término en el texto y 0 si no aparece, los vectores quedarían como sigue:

Texto1= [1 1 0 0 1 0 1 1 1 1 1 0 0 0]

Texto2= [1 0 1 0 0 0 0 0 0 0 0 1 1 0]

La fórmula del coseno para calcular la distancia entre los vectores es la siguiente:

[2] Tomado de (IEEE, 2014)

Revista Cubana de Ciencias Informáticas
ISSN: 2227-1899 | RNPS: 2301
Vol. 8, No. 4, Octubre-Diciembre, 2014
Págs. 130-144
http://rcci.uci.cu

$$\cos(Vx, Vy) = \frac{Vx * Vy}{|Vx| * |Vy|} \quad (1)$$

Donde:

Vx y Vy son los vectores de los textos 1 y 2 respectivamente.

Máxima subsecuencia común

Un aspecto importante a considerar cuando se trata el tema de la detección de plagio es mencionado por (Elizalde, 2011). Se refiere a que al buscar plagio se prefieren cadenas largas ya que a mayor longitud, mayor es la probabilidad de que el fragmento sea producto de una copia y no de una coincidencia casual. A partir de este concepto surge el algoritmo *Longest Common Subsequence* (LCS por sus siglas en inglés). El mismo es capaz de devolver el total de palabras coincidentes en cada sentencia u oración de los textos. Este algoritmo es utilizado para la comparación de textos mediante la herramienta Diff de Unix, que básicamente comprueba las diferencias entre dos versiones de un mismo archivo, muy común en las herramientas de versionado. Su aplicación cubre el rasgo (2) de los antes mencionados. La Figura 1 muestra la máxima subsecuencia común en dos frases similares.

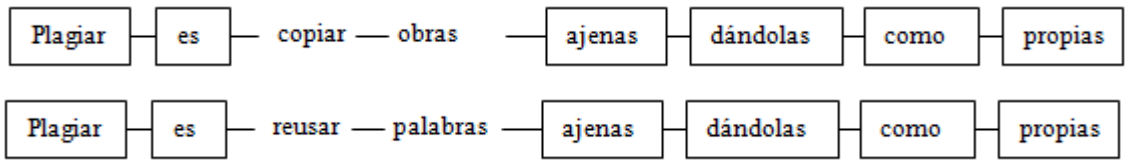

Figura 1. Máxima subsecuencia común entre dos frases similares.

Técnica de aprendizaje automático para la clasificación

Una vez empleados los algoritmos descritos para la detección de las coincidencias en el texto, se propone la utilización de una técnica de aprendizaje automático. Esta técnica permitirá determinar si un texto es plagiado o no a partir de los resultados que arroje la aplicación de cada uno de los métodos anteriores. Dentro de la rama de la inteligencia artificial existen varios métodos de aprendizaje automático para la clasificación (Chong, 2013) que permiten la realización de esta tarea. Luego de un estudio realizado se decidió explotar el algoritmo de clasificación Naïve Bayes o clasificador bayesiano ingenuo.

Este algoritmo es ampliamente usado en procesos de clasificación. Se le considera como una forma especial, o como el modelo más simple de clasificación basado en una Red Bayesiana (Hernández et al., 2004). Es utilizado para predecir

Revista Cubana de Ciencias Informáticas
ISSN: 2227-1899 | RNPS: 2301
Vol. 8, No. 4, Octubre-Diciembre, 2014
Págs. 130-144
http://rcci.uci.cu

la clase a la que pertenece una instancia determinada, suponiendo que las características de dicha instancia son independientes (Inza et al., 2000), como se muestra en la Figura 2.

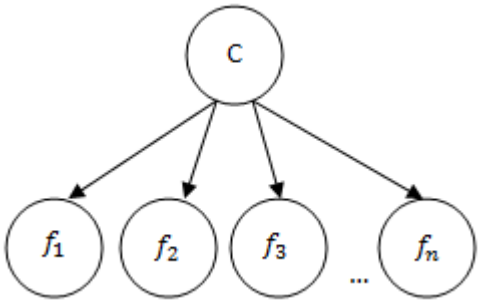

Figura 2. Estructura de la red Naïve Bayes. Fuente: Elaboración propia basado en (Friedman et al., 1997)

Para el caso que se aplica, se toman como las "características" el resultado de cada uno de los algoritmos de detección de plagio, con el fin de clasificar un texto en las siguientes clases "Plagiado" o "No plagiado".

Con relación al cálculo de la probabilidad de una hipótesis (Mitchell, 1997) define que, dado un número de características $\{f_1 \dots f_n\}$ conocidas, para un conjunto de entrenamiento de referencia, el clasificador Naïve Bayes plantea que:

$$P_c = argmaxp_c(C = c) * \prod_{i=1}^{n} p(F_i = f_i | C = c) \quad (2)$$

Para confeccionar la propuesta de solución se implementaron los algoritmos de detección de coincidencias antes presentados. Lo que posibilitó tener varios criterios para un documento que entra en revisión y definir de manera acertada cuándo el documento analizado posee coincidencias con otros ya publicados o en proceso. A continuación, la Figura 3 muestra un flujo de la propuesta de solución.

Revista Cubana de Ciencias Informáticas
ISSN: 2227-1899 | RNPS: 2301
Vol. 8, No. 4, Octubre-Diciembre, 2014
Págs. 130-144
http://rcci.uci.cu

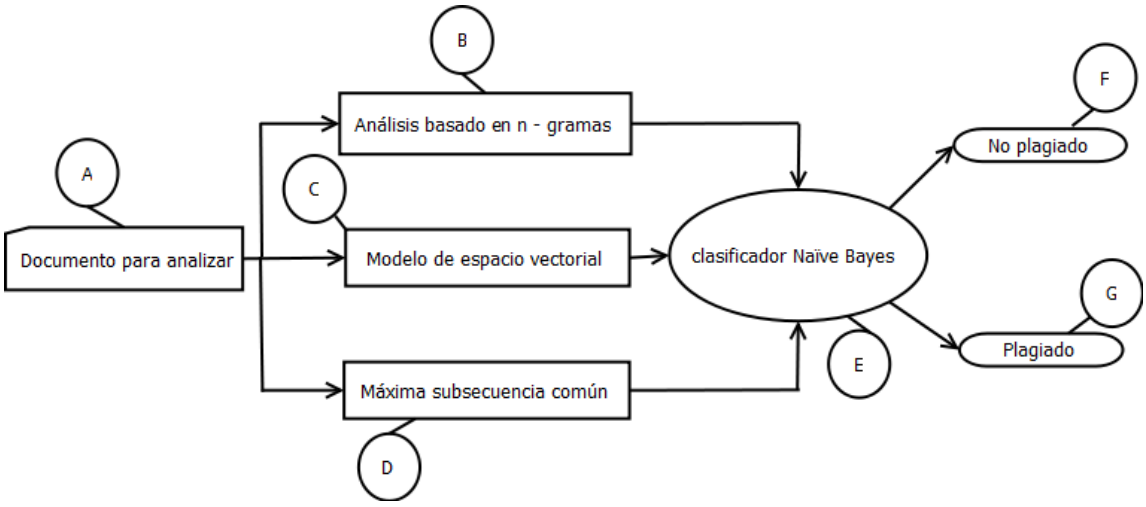

Figura 3. Representación del flujo de la propuesta de solución.

A: Documento para analizar y definir el nivel de coincidencias con los ya publicados o en proceso de revisión.

B, C, D: Algoritmos que procesan el documento de consulta en busca de similitudes con otros documentos y en función de sus resultados informan en qué porciento el documento es plagiado.

E: El Clasificador Naïve Bayes, tiene como entrada los valores resultantes de los algoritmos aplicados en la fase anterior. Con estos valores se ejecuta el algoritmo y su resultado definirá si el documento finalmente es plagiado o no.

F: No plagiado es uno de los tipos de clase para la clasificación. Este tipo afirma que no existe plagio en el documento especificado en el comienzo del flujo.

G: Plagiado es el otro tipo de clase para la clasificación. Este tipo afirma la existencia de plagio en el documento que se está analizando.

Resultados y discusión

Aplicación de la técnica de clasificación Naïve Bayes

Con el objetivo de lograr un mayor entendimiento del procedimiento propuesto se ilustra un ejemplo de cómo debe ser su funcionamiento básico. Para ello se supone que se tiene un conjunto de entrenamiento *E* con un total de 300instancias de documentos: 210 intencionalmente plagiados y 90 no plagiados. Además, se tienen dos categorías para representar el resultado de la aplicación de cada uno de los algoritmos de detección de plagio:

- Bajo: Si como resultado del algoritmo se obtiene que el texto coincida con otros en menos de un 30%.
- Alto: Si como resultado del algoritmo se obtiene que el texto coincida con otros en más de un 30%.

Revista Cubana de Ciencias Informáticas
ISSN: 2227-1899 | RNPS: 2301
Vol. 8, No. 4, Octubre-Diciembre, 2014
Págs. 130-144
http://rcci.uci.cu

En la Tabla 1 se muestran la cantidad de documentos por categoría según el resultado de la aplicación de los algoritmos, ya sea el análisis basado en n – gramas de palabras, modelo de espacio vectorial o máxima subsecuencia común. Estos datos son utilizados posteriormente para la clasificación.

Tabla 1. Datos del conjunto de entrenamiento para la clasificación.

Clasificación/Resultado de algoritmo	**Análisis basado en n - gramas**		**Modelo de espacio vectorial**		**Máxima subsecuencia común**	
	Bajo	**Alto**	**Bajo**	**Alto**	**Bajo**	**Alto**
Plagiados	28	182	48	162	87	123
No Plagiados	90	0	60	30	59	31

Se quiere comprobar si un texto es plagiado sabiendo que el resultado de la aplicación de los métodos de detección de plagio es el siguiente:

- Análisis basado en n-gramas de palabras: 12,1% (Bajo).
- Modelo de espacio vectorial: 46,7% (Alto).
- Mayor subsecuencia común: 32,3% (Alto).

Paso 1: Calcular las probabilidades a priori: Son las probabilidades de que ocurra una clase u otra, calculadas a partir de los datos obtenidos del conjunto de entrenamiento.

$$P_{(C=c)} = \frac{Cantidad_c}{Total} \tag{3}$$

$$P_{(C=Plagiado)} = \frac{210}{300} = 0{,}7$$

$$P_{(C=NoPlagiado)} = \frac{90}{300} = 0{,}3$$

Paso 2: Calcular las probabilidades condicionales: Probabilidades referentes a la ocurrencia de un evento dado otro. Al igual que las probabilidades a priori, se calculan a partir de los datos obtenidos del conjunto de entrenamiento y los datos mostrados en la Tabla 1.

$$P_{(n-gramas=bajo \mid C=Plagiado)} = \frac{28}{210} = 0{,}13$$

Revista Cubana de Ciencias Informáticas
ISSN: 2227-1899 | RNPS: 2301
Vol. 8, No. 4, Octubre-Diciembre, 2014
Págs. 130-144
http://rcci.uci.cu

$$P_{(n-gramas=bajo \mid C=NoPlagiado)} = \frac{90}{90} = 1$$

$$P_{(modeloespaciovectorial=alto \mid C=Plagiado)} = \frac{162}{210} = 0{,}77$$

$$P_{(modeloespaciovectorial=alto \mid C=NoPlagiado)} = \frac{30}{90} = 0{,}33$$

$$P_{(máximasubsecuenciacomún=alto \mid C=Plagiado)} = \frac{123}{210} = 0{,}59$$

$$P_{(máximasubsecuenciacomún=alto \mid C=NoPlagiado)} = \frac{31}{90} = 0{,}34$$

Paso 3: Calcular la probabilidad a posteriori. Se utilizan las probabilidades obtenidas de los dos pasos anteriores para aplicar la fórmula (2).

$$P_{c=\,plagiado} = 0{,}7 * \ 0{,}13 * 0{,}77 * 0{,}59 = 0{,}04$$

$$P_{c=\,noplagiado} = 0{,}3 * \ 1 * 0{,}33 * 0{,}34 = 0{,}03$$

Como resultado final se tiene que $0{,}04 > \ 0{,}03$ por tanto se clasifica el texto como “Plagiado”.

Finalmente, en la comprobación de los resultados fue utilizada la herramienta WEKA (acrónimo de *Waikato Environment for Knowledge Analysis*). Es una herramienta desarrollada por la Universidad de Waikato en Nueva Zelanda. Incluye una colección de herramientas para el procesado de datos y un conjunto de algoritmos de aprendizaje automático para la experimentación y análisis (Witten et al., 2011). Provee un paquete de algoritmos para la clasificación, entre ellos el Naïve Bayes. En la Tabla 2 se visualizan los resultados obtenidos en la ejecución del algoritmo Naïve Bayes utilizando un *dataset* con 300 instancias.

Tabla 2. Resultados de la ejecución del algoritmo Naïve Bayes en la herramienta WEKA.

Parámetro	**Valor**
Porciento de instancias correctamente clasificadas	97.67 %
Porciento de instancias incorrectamente clasificadas	2.34 %
Promedio de verdaderos positivos (TP)	0.98
Promedio de falsos positivos (FP)	0.01

Promedio de cobertura (Recall)[3]	0.98
Promedio de precisión (Precision)[4]	0.98
Clasificados como Plagiados	203
Clasificados como No plagiados	90
Clasificados incorrectamente	7
Número total de instancias	300

La Figura 4 muestra la matriz de confusión que retorna la herramienta WEKA. En ella se puede visualizar el resultado de la clasificación. Los valores de la diagonal corresponden a las instancias correctamente clasificadas y el resto a los errores. De los 210 documentos con plagio 203 fueron correctamente clasificados y 7 con error. De los 90 documentos no plagiados, todos fueron correctamente clasificados.

```
=== Confusion Matrix ===

   a   b   <-- classified as
 203   7 |   a = plagiado
   0  90 |   b = no_plagiado
```

Figura 4. Matriz de confusión.

En la Figura 5 se muestra una representación gráfica de los errores ocurridos en el proceso de clasificación con el algoritmo Naïve Bayes. El color azul identifica a las instancias clasificadas en la clase “Plagiados” y el color rojo representa a las instancias clasificadas en la clase “No plagiados”. Las cruces son las clasificadas correctamente y los cuadrados las clasificadas incorrectamente. Se puede obtener la representación para ver en cuál atributo se comete más error y en cuál menos. Por ejemplo las gráficas (a), (b) y (c) de la Figura 5 representan los errores teniendo en cuenta las tres características utilizadas para la clasificación: (a) el análisis basado en n-gramas, (b) el análisis basado en el modelo de espacio vectorial y (c) el análisis de la máxima subsecuencia común. La gráfica (d) muestra la relación entre el resultado que se predecía de la clasificación y el resultado real que se obtuvo.

[3] Mide la proporción de términos correctamente reconocidos respecto al total de términos reales.

[4] Mide el número de términos correctamente reconocidos respecto al total de términos predichos, sean estos verdaderos o falsos términos.

Revista Cubana de Ciencias Informáticas
ISSN: 2227-1899 | RNPS: 2301
Vol. 8, No. 4, Octubre-Diciembre, 2014
Págs. 130-144
http://rcci.uci.cu

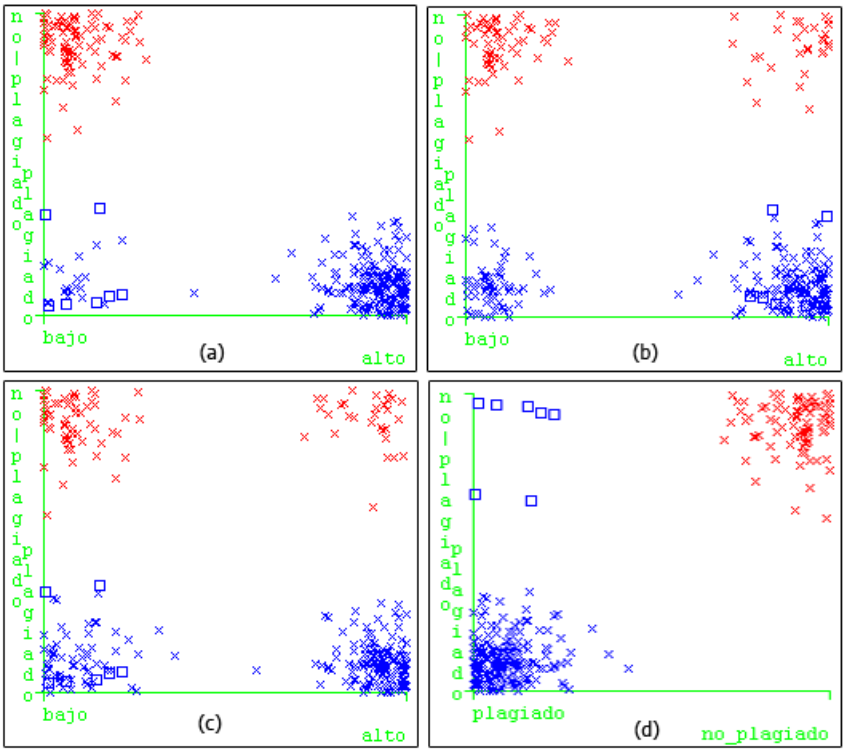

Figura 5. Errores en la clasificación, teniendo en cuenta la relación entre: (a) el atributo n-gramas y la clasificación (b) el atributo modelo de espacio vectorial y la clasificación (c) el atributo máxima subsecuencia común y la clasificación (d) la clasificación real y la predicción.

Conclusiones

A partir del estudio realizado, los autores de este reporte consideran que el uso de herramientas o métodos para detectar coincidencias de texto en documentos digitales, ha aumentado considerablemente en los últimos años. Teniendo en cuenta esta premisa, se considera que el principal aporte de la presente investigación es la utilización de la técnica de clasificación Naïve Bayes, permitiendo de esta forma procesar la salida de los métodos de detección de coincidencias aplicados. De esta manera, se tiene la posibilidad de combinar el resultado dado por varios algoritmos, lo que permite tener en cuenta diferentes criterios o conceptos de la detección de plagio. Por lo tanto, se puede afirmar que con este enfoque se cubre más de un rasgo en el proceso de análisis de los documentos.

La aplicación de la técnica propuesta, apoya la detección de coincidencias entre documentos de texto llevada a cabo como parte de los procesos de revisión de artículos. Lo anterior, evita que los documentos a publicar tengan más de un

30% de similitud con textos publicados previamente, elevando la calidad y fiabilidad del proceso de revisión. La posibilidad de detectar plagio desde una herramienta web local, soluciona el problema del consumo de ancho de banda, al no tener que utilizar herramientas en línea para realizar esta tarea. Además, se elimina la incertidumbre del manejo real que dan las herramientas en línea a la información que aún no ha sido publicada, garantizando de esta forma la integridad de las publicaciones.

La utilización de esta técnica puede generalizarse, y ser aplicada en la identificación de contribuciones que sean enviadas a más de una revista o evento a la vez. Lo anterior contribuye a reducir la posibilidad de que se publique un mismo artículo en varios escenarios, lo que denotará originalidad e innovación en los manuscritos que logren ser aceptados. De igual manera influye positivamente en el nivel y calidad científica de los documentos que sean sometidos a un proceso de arbitraje. Se recomienda la inclusión en la primera etapa de otros métodos de detección de coincidencias de texto, además de los propuestos, lo que aumentará la detección certera del plagio.

Referencias

- CEDEÑO, L. A. B. Detección automática de plagio en texto. Tesis desarrollada dentro del Máster en Inteligencia Artificial, Reconocimiento de Formas e Imagen Digital Universidad Politécnica de Valencia, 2008.
- CLOUGH, P. Old and new challenges in automatic plagiarism detection. National UK Plagiarism Advisory Service, 2003, 76.
- CHONG, M. Y. M. A Study on Plagiarism Detection and Plagiarism Direction Identification Using Natural Language Processing Techniques. A thesis submitted in partial fullment of the requirements of the University of Wolverhampton for the degree of Doctor of Philosophy Research Group in Computational Linguistics University of Wolverhampton, UK, 2013.
- ELIZALDE, V. Estudio y desarrollo de nuevos algoritmos de detección de plagio. Tesis de Licenciatura en Ciencias de la Computación Departamento de Computación, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, 2011.
- FRIEDMAN, N., D. GEIGER AND M. GOLDSZMIDT. Bayesian Network Classifiers. In *Machine Learning*. Netherlands: Kluwer Academic Publishers, 1997, vol. 29, p. 131–163.
- HERNÁNDEZ, J., M. J. RAMÍREZ AND C. FERRI *Introducción a la Minería de Datos*. Edtion ed. Madrid: Pearson Prentice Hall, 2004.

- IEEE. Institute of Electrical and Electronics Engineers. A plagiarism FAQ. 2014, [cited 21 de marzo 2014]. Available from Internet:<http://www.ieee.org/web/publications/rights/plagiarismFAQ.htm>.
- INZA, I., P. LARRAÑAGA, R. ETXEBERRIA AND B. SIERRA. Feature Subset Selection by Bayesian network-based optimization. In *Artificial Intelligence.* Elsevier Science, 2000, vol. 123, p. 157–184.
- MANCHEGO, F. E. A. Sistema de información de detección de plagio en documentos digitales usando el método Document Fingerprinting. Tesis para optar por el Título de Ingeniero Informático Pontificia Universidad Católica del Perú, 2010.
- MITCHELL, T. M. *Machine Learning*. Edtion ed.: McGraw-Hill Science/Engineering/Math, 1997.
- PINTO, D., D. VILARIÑO, C. BALDERAS, M. TOVAR, et al. Evaluating n-gram Models for a Bilingual Word Sense Disambiguation Task. Computación y Sistemas, 2011, 5(2), 209-220.
- RAE. Diccionario de la Lengua Española. 2014, [cited 18 de marzo 2014]. Available from Internet:<http://www.rae.es/rae.html>.
- STAMATATOS, E. Plagiarism detection using stopword n-grams. Journal of the American Society for Information Science and Technology, 2011, 62(12), 2512–2527.
- VEGA, J. F. S. Detección automática de plagio basada en la distinción y fragmentación del texto reutilizado. . Tesis sometida como requisito parcial para obtener el grado de: Maestro en Ciencias en el Área de Ciencias Computacionales Instituto Nacional de Astrofísica, Óptica y Electrónica, 2011.
- WITTEN, I. H., E. FRANK AND M. A. HALL *Data Mining Practical Machine Learning Tools and Techniques*. Edtion ed.: Morgan Kaufmann Publishers, 2011. ISBN 978-0-12-374856-0.
- ZECHNER, M., M. MUHR, R. KERN AND M. GRANITZER. External and Intrinsic Plagiarism Detection using Vector Space Models. In *3rd Pan Workshop. Uncovering plagiarism, authorship and social software Misuse. 25th Annual Conference of the spanish society for natural language processing.* SEPLN, 2009, p. 47-55.