RUSC Revista de Universidad y Sociedad del Conocimiento
Universities and Knowledge Society Journal
http://rusc.uoc.edu

ARTÍCULO

Determinación de estilos de escritura para la detección de similitudes entre documentos digitales

Yohandri Ril Gil
rilltt@uci.cu
Asesor de Tecnologías y Seguridad Informática del Centro Tecnologías para la Formación de la Universidad de las Ciencias Informáticas

Yuniet del Carmen Toll Palma
ytoll@uci.cu
Asesora de Calidad del Centro Tecnologías para la Formación de la Universidad de las Ciencias Informáticas

Eddy Fonseca Lahens
elahens@uci.cu
Especialista del Centro IdeoInformática,
Universidad de las Ciencias Informáticas

Fecha de presentación: marzo de 2013
Fecha de aceptación: abril de 2013
Fecha de publicación: enero de 2014

Cita recomendada
Ril, Y., Toll, Y.C. y Fonseca, E. (2014). Determinación de estilos de escritura para la detección de similitudes entre documentos digitales. *Revista de Universidad y Sociedad del Conocimiento (RUSC), 11*(1). doi http://dx.doi.org/10.7238/rusc.v11i1.1783

Resumen

Todo lo inherente al intelecto humano es susceptible de actos de plagio: obras científicas y literarias tales como artículos, tesis, obras audiovisuales, planos y proyectos, códigos fuentes de programas, entre otros. Sin embargo, el presente trabajo dedica especial atención a la existencia de este fenómeno en obras escritas, en concreto documentos digitales provenientes de lenguajes naturales o de programación, y centra su objetivo en el desarrollo y aplicación de un modelo matemático que permite determinar el estilo de escritura empleado en la redacción de los textos. Los resultados que se esperan obtener a partir de la aplicación del procedimiento servirán de base para la reducción en el número de documentos que se deben comparar en el análisis y detección de similitudes entre estos documentos. De forma experimental se aplica el procedimiento a un grupo de artículos clasificados por temáticas y autores y que difieren entre ellos en el estilo de escritura utilizado para su redacción.

Palabras clave

estilo de escritura, documentos digitales, plagio, procedimiento

Determination of writing styles to detect similarities in digital documents

Abstract

Anything involving human intellect is at risk of being plagiarised. This includes scientific and literary works such as articles, theses, audiovisual works, plans, projects and computer programs. However, this article pays special attention to the existence of this phenomenon in written works in general, and in digital documents in natural or programming languages in particular. The objective of the research is to develop and apply a mathematical model that allows the writing style used in the drafting of texts to be determined. The results obtained from the application of the procedure are intended to serve as the basis for reducing the number of documents that need to be compared in order to analyse and detect similarities in them. The procedure was experimentally applied to a set of articles classified by topic and author, where the writing styles used to draft them differed.

Keywords

writing style, digital documents, plagiarism, procedure

1. Introducción

Las ventajas aportadas por la aparición de las tecnologías de la información y la comunicación (TIC) son incuestionables, numerosos ejemplos lo demuestran. Sin embargo, varios elementos negativos las acompañan, por ejemplo el plagio, que sin duda es uno de los más representativos. La Real Academia Española define el término «plagiar» (Española, 2001) de forma sencilla y tajante:

Plagiar: copiar en lo sustancial obras ajenas, dándolas como propias.

Todo lo inherente al intelecto humano es susceptible de actos de plagio: obras científicas y literarias tales como artículos, tesis, cinematografías, partituras, obras audiovisuales, planos y proyectos,

códigos fuentes de programas, sitios webs, etc. Sin embargo, en el presente trabajo dedicamos especial atención a la existencia de este fenómeno en obras escritas, en concreto documentos digitales provenientes de lenguajes naturales o de programación.

Utilizar elementos presentes en otras obras resulta una práctica común. El modo de hacerlo define la intencionalidad a la hora de respetar o no la autoría de las fuentes. Entre los distintos métodos empleados se encuentran los siguientes:

- **Correcto:** se muestra de forma precisa la procedencia del material textual referenciado, añadiendo información acerca del autor de lo citado.
- **Parafraseo:** a partir de la interpretación, se exponen con palabras propias determinadas ideas presentes en otros documentos.
- **Múltiples fuentes:** se crea un nuevo documento a partir de la utilización textual o modificada de fragmentos provenientes de otros documentos.
- **Mosaico:** se desordenan segmentos de un documento original para ser empleados textualmente en un nuevo documento.
- **Modificaciones puntuales:** se utilizan segmentos de un documento original, intercalando palabras o frases para su modificación.
- **Copiar y pegar:** el documento es copiado de modo parcial o total y sin modificaciones. Es relativamente fácil de detectar partiendo de una comparación para determinar si existen o no diferencias entre los documentos.

Los avances tecnológicos han permitido contar con sistemas capaces de detectar el plagio entre documentos; sin embargo, resulta indispensable la supervisión humana en este proceso. Una acusación de plagio es grave por lo que no debe realizarse de forma irresponsable. La utilización de herramientas automatizadas garantiza una discriminación en el número de documentos en que debe intervenir el hombre, en ocasiones sobre la base de heurísticas como es el caso de determinar entre varias páginas webs similares que la original es aquella de mayor *ranking* de posicionamiento. En el caso de una comparación entre trabajos realizados por estudiantes, ante la similitud podría otorgársele la originalidad al alumno de mejor rendimiento. Lamentablemente estas heurísticas no siempre son válidas, ni siquiera puede asegurarse que lo sean en la mayoría de los casos. Como sucede con frecuencia, depende de una valoración subjetiva el análisis planteado, por lo que en adelante se empleará el término «determinación de similitudes» en lugar de «plagio».

Múltiples artículos abordan temáticas relacionadas con la determinación de similitud entre documentos. Entre ellos destacan los siguientes:

«Detección de similitud entre documentos usando términos relevantes» (Cooper *et al*. 2002)
Se realiza la comparación entre dos documentos mediante el procesamiento previo de ambos y la búsqueda de *tokens*[1] definidos con anterioridad, como pueden ser nombres propios, siglas, localiza-

1. Un *token*, también llamado componente léxico, es una cadena de caracteres que tiene un significado coherente en cierto lenguaje natural o de programación.

ciones, abreviaciones, etc. A cada término se le asigna un coeficiente IQ, equivalente a la cantidad de información que aporta su aparición en el texto. La puntuación de similitud entre dos textos refleja proporcionalidad con el número de términos presentes en uno y no en el otro.

«Herramienta de soporte para la detección de plagio en documentos de texto» (Gruner y Naven, 2005)

Se muestra un método para el análisis del estilo de escritura a partir de un comportamiento estadístico. Se busca determinar el modo de utilización del lenguaje por el autor y compararlo con otros documentos. La búsqueda de estilos puede realizarse por párrafos o por fragmentos del texto.

«Check: un sistema para la detección de plagio en documentos» (Si *et al.*, 1997)

Determina la similitud basándose en la frecuencia de aparición de los términos. Se genera un vector de pesos para cada documento comparado, y el coseno entre estos vectores es empleado como parámetro en una función que devuelve la estimación de la similitud. El procedimiento se repite para cada sección de los documentos hasta determinar la existencia o no de copia entre ellos.

«Sim: una herramienta para detectar similitudes entre códigos de programas» (Gitchell y Tran, 1999)

El artículo trata sobre detección de similitudes entre códigos fuentes. Se separan los códigos en *tokens* para luego calcular un puntaje de alineación entre las dos cadenas de *tokens*. A cada falla o reemplazo en la alineación le es asignado un peso. La similitud se obtiene mediante la expresión:

$$s = \frac{2 \times score(p1,p2)}{score(p1,p2) + score(p1,p2)} \qquad [I]$$

Donde **score (p1, p2)** representa el puntaje de alineación entre el código 1 y el 2.

«Plagio en lenguaje natural y de programación; un acercamiento a las herramientas y tecnologías actuales» (Clough, 2000)

Resume un grupo de herramientas disponibles para la detección de similitudes entre textos provenientes tanto de lenguajes naturales como de programación. Describe de manera sintetizada los elementos distintivos de algunos algoritmos empleados en dichas herramientas para la comparación como es el caso de la determinación de la subsecuencia común más larga.

El objetivo de la investigación es brindar un desarrollo e implementación de un modelo matemático que permite determinar el estilo de escritura empleado en la redacción de los textos.

Método de investigación o teoría utilizada

Para el desarrollo de la investigación se utilizaron varios métodos científicos. Por ejemplo, el método de revisión bibliográfica se aplicó con el objetivo de realizar una búsqueda sobre las fuentes de

información que se emplean para el sustento teórico del estudio. El método de análisis y síntesis se utilizó en todo el proceso investigativo, principalmente en la precisión de los fundamentos teóricos relacionados con la detección y similitudes de documentos digitales y el análisis e interpretación de los resultados obtenidos de la aplicación de la herramienta para detección de plagio en los documentos digitales. Además se utilizaron métodos del nivel estadístico como el método de estadística descriptiva, para el procesamiento de los datos obtenidos en la aplicación de herramientas para la detección de similitudes en los documentos digitales, y el método de estadística inferencial, para la toma de decisiones sobre el rechazo o no de los datos obtenidos durante el proceso de detección de similitudes en los documentos digitales.

2. Desarrollo

El mayor reto de los sistemas para la detección de similitudes entre documentos lo constituye el gigantesco volumen de información que hay que procesar; es por ello por lo que se requiere una clasificación previa del conjunto de muestra o documentos originales con los que se cuenta para la comparación.

Comúnmente se tienen en cuenta para la clasificación criterios como tipo de documento, idioma, categoría, subcategoría, palabras clave, autores, fechas, entre otros. Nuestra propuesta consiste en incorporar el estilo de escritura como criterio comparativo a la hora de determinar la originalidad de un documento. De esta manera, el número de documentos que se van a comparar en la búsqueda de similitudes se reduciría a aquellos que guarden relación directa en lo que a clasificación se refiere con el documento comparado, y que a su vez posean un estilo de escritura similar.

2.1. Procedimiento para la extracción del estilo de escritura

Todo texto puede ser representado a través de un modelo estadístico que identifique sus características intrínsecas, las cuales se relacionan con el estilo de escritura del autor. Entre ellas podemos mencionar:

Palabras de paro: se refiere a la utilización de artículos, adverbios, preposiciones y conjunciones. La frecuencia en el uso de estas palabras denota un estilo en el modo de escribir del autor. Consecuentemente, se define el promedio de palabras de paro como:

$$Pp = \frac{Cpp}{Tp} * 100\ \% \qquad [II]$$

Donde *Cpp* indica cuántas palabras de paro fueron utilizadas y *Tp* el total de palabras del texto.

Nivel de dificultad: determina el grado de educación requerido por una persona para la comprensión del texto. Existen varios índices destinados a la obtención de este nivel. Gunning (Wikipedia,

2011), Dale y Chall (1948) y Flesch-Kincaid (DuBay, 2004) son algunos de ellos, si bien este último es uno de los más documentados y referenciados.

La expresión para determinar el índice de Flesch-Kincaid es como sigue:

$$I_{FK} = 1.599\Lambda - 1.015\beta - 31.517 \qquad [III]$$

Donde λ es el promedio por cada cien palabras de aquellas que tengan una sola sílaba y β corresponde a la longitud promedio de las oraciones medidas en número de palabras.

Variedad de vocabulario: propuesta por Honore (1979) intenta determinar la variedad en el vocabulario del autor en función del total de palabras no repetidas empleadas en el texto. Se utiliza para ello la expresión:

$$R = \frac{100\ log\ (M)}{M^2} \qquad [IV]$$

Donde *M* corresponde al número de palabras diferentes del texto.

En dependencia del tipo de documento analizado, reviste mayor o menor validez el cálculo de *R*. Un ejemplo claro lo constituyen documentos que por su naturaleza requieran de la constante repetición de palabras, como lo suelen ser determinados artículos especializados o los propios códigos fuentes de programas.

A raíz de lo anterior se introduce una aproximación propuesta por Yule (1944) como alternativa de cálculo, definiendo:

$$K = \frac{10^4 (\sum_i^{\infty} = 1\ i^2\ V_i - M)}{M^2} \qquad [V]$$

Donde V_i es el número de palabras que aparece *i* veces en el texto. *M* tiene el mismo significado que en el cálculo anterior.

Longitud promedio de oraciones: es una medida fiel de los conocimientos gramaticales empleados por el autor en la composición de sus oraciones.

$$Lp_0 = \frac{\sum_i^{No} = 1\ lo_i}{No} = \frac{Tp}{No} \qquad [VI]$$

Donde lo_i representa la longitud en palabras de la oración i-ésima, *No* identifica el número total de oraciones y *Tp* el total de palabras que componen el texto.

Longitud promedio de palabras: este término se encuentra vinculado directamente con la riqueza de vocabulario del autor y mide la facilidad que tiene para la utilización de palabras complejas[2].

$$Lp_p = \frac{Tc}{Tp} \qquad [VII]$$

Donde *Tc* se refiere al total de caracteres empleados, exceptuando el espaciado, y *Tp* identifica el total de palabras empleadas.

Finalmente se define el vector de estilo de escritura cuyos componentes acaban de ser descritos.

$$E < Pp, I_{FK}, K, Lp_0, Lp_p > \qquad [VIII]$$

Si bien es cierto que la determinación del estilo de escritura de un autor encierra en sí misma un grado de incertidumbre, conocer dicho grado permite establecer el margen de desviación adecuado a la hora de determinar el creador de un documento. Dentro de los principales parámetros que influyen en la incertidumbre pueden mencionarse la temática abordada, la extensión del documento en cuanto al número de párrafos, oraciones o palabras empleadas, la experiencia del autor, el destinatario del documento, entre otros.

La importancia de contar con la estimación estadística del modo de escritura radica en poseer un criterio comparativo para la selección y clasificación de documentos y determinar su autoría. Una de las aplicaciones la constituye la determinación de similitudes entre documentos, donde se requiere de la optimización del tiempo de procesamiento. Son gigantescas las bases de datos que alojan el material que se debe comparar, por lo que procesarlo todo no es una opción, ni siquiera lo es el procesar únicamente aquellos documentos que pertenezcan a una misma categoría de clasificación. Se necesita de otros argumentos que discriminen y reduzcan considerablemente el número de documentos que se van a procesar, como puede ser el empleo de un identificador del estilo de escritura.

3. Análisis de la investigación y resultados

Para aplicar la propuesta de determinación del estilo de escritura, fueron seleccionados 37 documentos pertenecientes a 5 autores y enmarcados en 4 áreas temáticas: educación, historia, medicina y cuentos infantiles. Dos de los autores coinciden en la temática de medicina. La figura 1 muestra la distribución de documentos por área temática.

2. Se consideran palabras complejas aquellas formadas por tres o más sílabas y que no representan nombres propios, prefijos, sufijos o palabras compuestas.

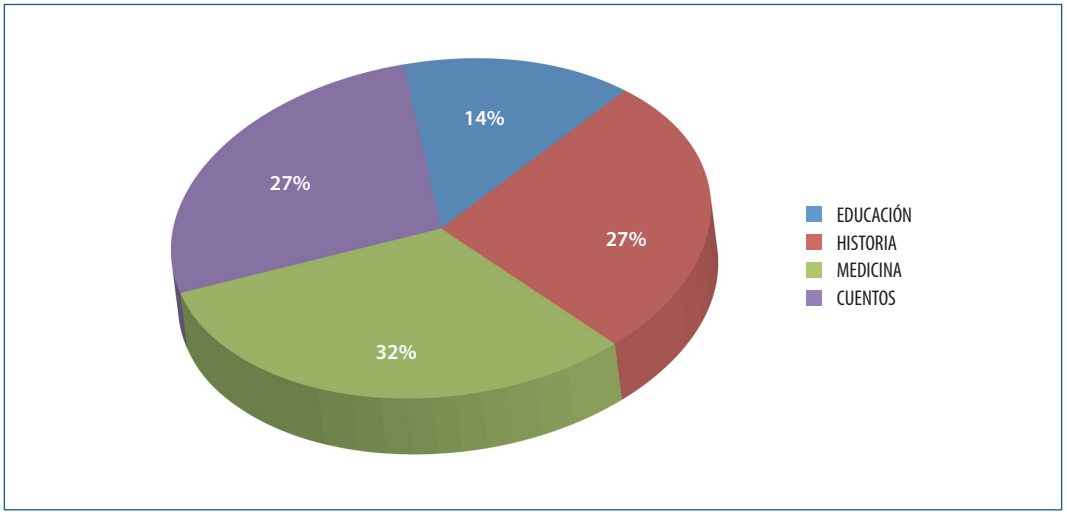

Figura 1. Distribución de documentos por área temática

Para determinar el vector de estilo de escritura de forma automatizada, se desarrolló una herramienta (figura 2) que permite la extracción de un grupo de estadísticas de los textos analizados como paso previo al cálculo del vector.

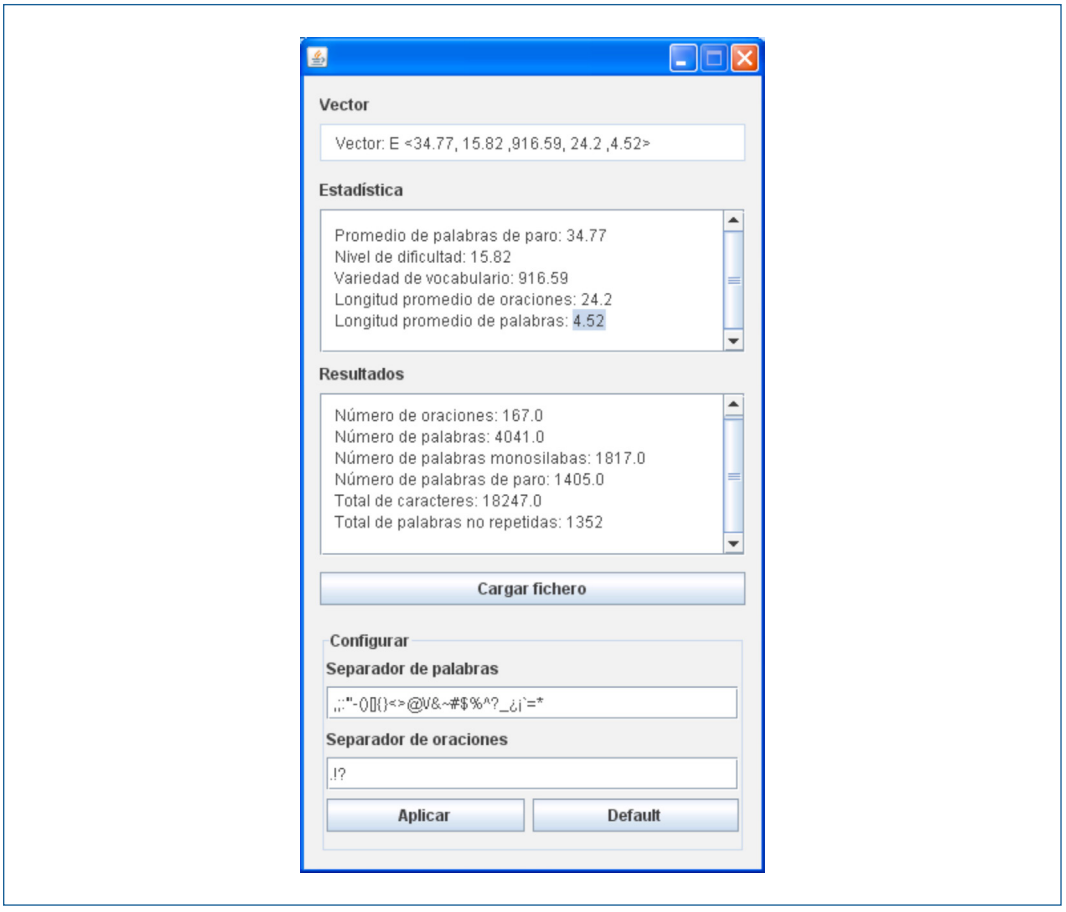

Figura 2. Herramienta para el cálculo del vector de estilo

A continuación se relacionan las características extraídas de los documentos:

- Total de oraciones
- Total de palabras
- Total de palabras monosílabas
- Total de palabras de paro
- Total de caracteres
- Total de palabras no repetidas
- Promedio de palabras de paro
- Nivel de dificultad
- Variedad de vocabulario
- Longitud promedio de oraciones (medida en palabras)
- Longitud promedio de palabras (medida en caracteres)

Para la determinación de tendencias y otras estadísticas se empleó una hoja de cálculo de Microsoft Office 2007. La tabla 1 muestra, para cada autor, los valores promedios alcanzados por los componentes del vector estilo. Como se puede apreciar, el parámetro que aporta la mayor diferenciación es el de variedad de vocabulario, y resaltan en este caso los autores 1 y 2, de la temática educación e historia respectivamente. Precisamente el autor 2, por el hecho de utilizar documentos narrativos de fácil comprensión, presenta el menor nivel de dificultad.

Tabla 1. Vector de estilo promedio por autor

	Promedio Pp	*Nivel Dificultad*	*V. Vocabulario*	*L.P. Oraciones*	*L.P. Palabras*
Autor 1	37,49	21,89	2482,43	20,46	5,314
Autor 2	35,88	11,40	2357,63	30,39	4,742
Autor 3	30,52	26,88	1011,01	12,41	5,162
Autor 4	24,21	28,96	1121,34	10,57	5,356
Autor 5	33,34	26,70	665,76	12,54	4,399

Resulta imprescindible determinar para cada autor cómo varían los parámetros del estilo de escritura, en otras palabras, cómo un mismo autor puede mantener o no su estilo de escritura a lo largo de un grupo de documentos que aborden una misma temática. En la tabla 2 se realiza un análisis por cada parámetro del vector para calcular su desviación promedio. Como podía preverse de la tabla anterior es precisamente la variedad de vocabulario el parámetro de mayor fluctuación como lo es la longitud promedio de palabra el de menor, tanto así que fue necesario incrementar a tres (3) lugares después de la coma la observación.

Tabla 2. Desviaciones promedio por cada parámetro del vector

	Promedio Pp	Nivel Dificultad	V. Vocabulario	L.P. Oraciones	L.P. Palabras
Autor 1	0,65	0,82	107,03	1,26	0,079
Autor 2	0,53	0,74	82,57	1,10	0,075
Autor 3	1,96	1,05	69,78	1,05	0,066
Autor 4	1,25	1,53	229,08	0,99	0,092
Autor 5	1,10	5,43	75,88	1,87	0,079

Cuando se realiza una revisión detallada de los documentos analizados, puede interpretarse que la variedad de vocabulario extraída guarda estrecha relación no solo con la riqueza en el uso del lenguaje del autor sino con la temática abordada y el público a quien se dirige.

Con el uso de la propia tabla de resultados, se infiere que aquellos autores con menores fluctuaciones en su estilo de escritura de un documento a otro serán precisamente quienes conserven con mayor énfasis una forma auténtica a la hora de redactar.

La figura 3 muestra de manera gráfica, a través de la representación de áreas, la distribución de desviaciones. Puede notarse que al autor 2 le corresponde el estilo de escritura más estable pues el área que le representa (color rojo) es la más homogénea. Todo lo contrario sucede con los autores 4 y 5, quienes presentan varios picos en sus respectivas áreas de distribución de desviaciones.

A pesar de que los autores 3 y 4 pertenecen ambos al área temática de medicina, existe una notable diferencia entre los estilos de escritura de cada cual.

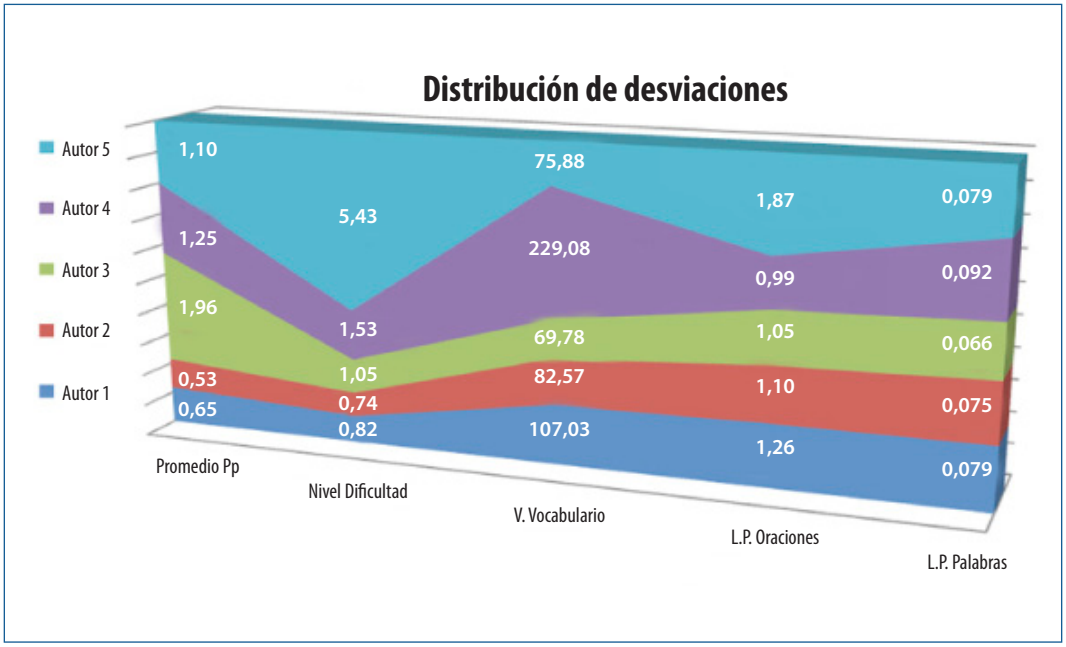

Figura 3. Áreas de distribución de desviaciones

El método propuesto para la determinación del estilo de escritura puede ser utilizado en un escenario donde se precise describir documentos cuya autoría haya sido validada. Es común la utilización de descriptores como autor, título, palabras clave, categoría, subcategoría, tipo de documento, entre otros; sin embargo, el contar con un descriptor como es estilo de escritura posibilitará discriminar significativamente el número de documentos que se deben comparar en la búsqueda de posibles plagios. En este sentido, cuando se pretenda analizar la presencia de plagio en un nuevo documento, solo será necesario comparar con un subconjunto conformado por aquellos de similar vector de estilo de escritura.

Aunque puede partirse de la concepción de que cada autor tiene su propio estilo de escritura, es importante considerar que las principales amenazas a la validez de este método radican en la selección de los documentos para establecer el estilo de escritura de un autor. Deben ser documentos cuya autoría sea debidamente demostrada, téngase en consideración los casos en que participa más de un autor en la redacción del documento. Cuando el documento no pertenece a las temáticas tradicionalmente abordadas por el autor, es aceptable que los índices de desviación varíen en un rango razonable, por lo que se recomienda emplear documentos de temáticas comúnmente abordadas por el autor. Finalmente y no menos importante, existen elementos como el propio proceso de aprendizaje y la realidad cambiante que rodea al individuo que van provocando paulatinamente variaciones en el estilo de escritura de cada autor. En este sentido, tendrá mayor fiabilidad la determinación del estilo de escritura cuanta mayor madurez y experiencia tenga el individuo desde el punto de vista de la redacción.

4. Discusión y conclusiones

Si bien el procedimiento propuesto se enfoca a la determinación de estilos de escritura, facilita a la comunidad académica la observación de similitudes entre documentos. Sin embargo, no constituye una solución definitiva al grave problema del plagio académico. Se hace necesaria la formación consciente y sistemática de valores como la responsabilidad y la honestidad.

¿Por qué insistir en la honestidad académica? Si no hay honestidad académica, se estará proyectando una imagen de conocimiento que no corresponde a la realidad de lo que auténticamente está en la estructura cognitiva. El comportamiento deshonesto en el ámbito académico es un modo de engaño y, sobre todo, una forma de autoengaño. Erosiona, desde la base, el propósito educativo de nuestra actividad docente. El concepto de honestidad por su carácter ético podríamos suponer que se encuentra sobreentendido, sin embargo esta suposición no es del todo acertada. Actos de plagio son cometidos intencionadamente por el o los involucrados, o bien de manera incidental o inconsciente.

¿Cómo influir en el fortalecimiento de estos valores en nuestros estudiantes? Cuando mostramos compromiso y congruencia profesional, creando una atmósfera intelectual y pedagógica de alto nivel, logramos eliminar o disminuir la deshonestidad académica y sus efectos negativos. En el nivel de educación superior queda claro entonces que existen grandes posibilidades de aprendizaje y al

mismo tiempo grandes posibilidades de fraude. De cierta manera confiamos en que la mayoría de los estudiantes están conscientes de su obligación de aprender influenciados por la necesidad de obtener un título. Nuestra función educativa no puede estar limitada al empleo de herramientas de supervisión de plagio, esto no garantiza que los alumnos no actúen fraudulentamente. Debemos incidir en que la conducta ética esté basada en la libre aceptación de reglas de comportamiento que no se pueden imponer por la fuerza de una autoridad.

Desde las primeras edades el ser humano, a través de los diversos procesos de aprendizaje por los que transita, va enriqueciendo su vocabulario, incrementando sus vivencias, su preparación y moldeando su estilo de escritura. Es por ello por lo que se percibe la creación de un estilo de redacción propio como un proceso paulatino. Para futuros trabajos se pretende ahondar en procedimientos para la obtención de las curvas de desarrollo de los estilos de escritura, así como en la caracterización de estos estilos, sobre la base de los parámetros que lo describen y que han formado parte primordial de este trabajo.

La apropiación de obras ajenas como propias continuará y evolucionará al ritmo en que lo hace la tecnología, estar preparados para contrarrestar este fenómeno resulta de vital importancia. Gran parte de las investigaciones consultadas para esta investigación constituirán la base para el desarrollo de futuros trabajos en el área de la detección de similitudes entre documentos, y la propuesta presentada servirá como punto de partida en la determinación de qué documentos comparar. La extracción del vector de estilo marca la diferencia entre los autores, independientemente de si abordan o no la misma temática. Con la aplicación del modelo matemático de la propuesta a un grupo considerable de documentos pudo comprobarse que realmente existen tendencias a la hora redactar y el hacerlo significa incorporar un sello de autenticidad a lo escrito.

Bibliografía

Clough, P. (2000). Plagiarism in natural and programming languages: an overview of current tools and technologies. *Research Memoranda: CS-00-05, Department of Computer Science, University of Sheffield, UK,* 1-31. Consultado en: http://ir.shef.ac.uk/cloughie/papers/plagiarism2000.pdf

Cooper, J. W., A. R. Coden, et al. (2002). Detecting similar documents using salient terms. En: *Proceedings of the eleventh international conference on Information and knowledge management*. Nueva York: ACM. Consultado en: http://www.labsoftware.com/flahdo/Papers/CIKMDuplicates.pdf

Dale, E. and J. S. Chall (1948). A formula for predicting readability. *Educational Research Bulletin*, *27*(1), 11-20. Consultado en: http://www.ecy.wa.gov/quality/plaintalk/resources/classics.pdf

Dubay, W. H. (2004). *The principles of readability*. CA: Impact Information. Consultado en: http://files.eric.ed.gov/fulltext/ED490073.pdf

Española, R. A. (Ed.) (2001). *Diccionario de la Real Academia Española*. Madrid: Real Academia Española.

Gitchell, D.; Tran, N. (1999). Sim: a utility for detecting similarity in computer programs. En: *The proceedings of the thirtieth SIGCSE technical symposium on Computer Science Education*. Nueva York: ACM. Consultado en: http://www.eng.uwi.tt/depts/elec/staff/feisal/ee302/sim-gitchell.pdf

Gruner, S.; Naven, S. (2005). Tool support for plagiarism detection in text documents. En: *Proceedings of the 2005 ACM symposium on Applied computing*. Nueva York: ACM. Disponible en: http://dl.acm.org/citation.cfm?id=1066677.1066854. doi http://dx.doi.org/10.1145/1066677.1066854

Honoré, A. (1979). Some simple measures of richness of vocabulary. *Boletín de la Association for Literary and Linguistic Computing*, *7*(2).

Si, A.; Leong, H. V.; Lau, R. W. H. (1997). Check: a document plagiarism detection system. En: *Proceedings of the 1997 ACM symposium on Applied computing*. Nueva York: ACM. Consultado en: http://www.cs.cityu.edu.hk/~rynson/papers/sac97.pdf. doi http://dx.doi.org/10.1145/331697.335176

Wikipedia (2011). Gunning fog index. *Wikipedia*. Online: Wikipedia.org. Consultado en: http://en.wikipedia.org/wiki/Gunning_fog_index

Yule, G. U. (1944).The statistical study of literary vocabulary. Cambridge, Cambridge [Eng.] University Press. *Journal of the Royal Statistical Society volumen 107*(número 2), 129-131. Disponible en: http://www.jstor.org/discover/10.2307/2981280?uid=3737824&uid=2129&uid=2&uid=70&uid=4&sid=21102626763567. doi http://dx.doi.org/10.2307/2981280

Sobre los autores

Yohandri Ril Gil
rilltt@uci.cu
Asesor de Tecnologías y Seguridad Informática del Centro Tecnologías para la Formación de la Universidad de las Ciencias Informáticas

Yohandri Ril Gil es ingeniero en Telecomunicaciones y Electrónica y profesor asistente de la Universidad de las Ciencias Informáticas, especificamente de Teleinformática I y II en la carrera de Informática. Es asesor de Tecnologías y Seguridad Informática del Centro Tecnologías para la Formación. Está cursando la maestría de Educación a Distancia en la Universidad de las Ciencias Informáticas. Sus líneas de investigación fundamentales se centran en redes de telecomunicaciones, seguridad informática, plataformas virtuales de aprendizaje, calidad de objetos de aprendizaje y arquitectura de software. Ha publicado varios artículos en revistas de impacto como *No Solo Usabilidad* y *RUSC*, y en memorias de eventos tanto nacionales como internacionales.

Yuniet del Carmen Toll Palma
ytoll@uci.cu
Asesora de Calidad del Centro Tecnologías para la Formación de la Universidad de las Ciencias Informáticas

Yuniet del Carmen Toll Palma, máster en Calidad en software, es profesora asistente de la Universidad de las Ciencias Informáticas y colabora como especilista de información del Grupo de la Información y el Conocimiento del Centro Tecnologías para la Formación. Es responsable de la edición de las secciones «Producción de recursos didácticos» y «Archivo» del boletín *TeduScopio* del centro FORTES y editora general del boletín. Actualmente se desempeña como asesora de Calidad del Centro Tecnologías para la Formación. Sus líneas de investigación fundamentales se centran en la evaluación de la calidad de los objetos de aprendizaje y de otros productos de software, la gestión de la información y el conocimiento y la arquitectura de la información para los proyectos del centro FORTES. Ha publicado varios artículos en revistas de impacto como *EDUTEC*, *No Solo Usabilidad*, *REDC* y *RUSC*, y en memorias de eventos tanto nacionales como internacionales, en los que ha tenido una amplia participación.

Eddy Fonseca Lahens
elahens@uci.cu
Especialista del Centro IdeoInformática, Universidad de las Ciencias Informáticas

Eddy Fonseca Lahens es especialista del Centro IdeoInformática de la Universidad de las Ciencias Informáticas. Es jefe del proyecto Motor de categorización inteligente de contenidos y arquitecto en el proyecto Motor de categorización inteligente de contenido para correo electrónico. Su línea de investigación fundamental se centra en el desarrollo de soluciones informáticas para internet, el desarrollo de aplicaciones concurrentes y distribuidas, la categorización automática de contenidos y la automatización de la evaluación de los objetos de aprendizaje. Ha publicado varios artículos en revistas nacionales y en memorias de eventos tanto nacionales como internacionales.

Universidad de las Ciencias Informáticas
Carretera a San Antonio de los Baños, km 2 ½
Torrens, municipio de La Lisa
La Habana, Cuba