Procesamiento del Lenguaje Natural, Revista nº 45, septiembre 2010, pp 49-57 recibido 13-04-10 revisado 18-05-10 aceptado 20-05-10

Detección de plagio en documentos. Sistema externo monolingüe de altas prestaciones basado en n-gramas contextuales

Plagiarism detection in documents. High performance monolingual external analysis system based on contextual n-grams

Diego Antonio Rodríguez Torrejón
I.E.S. "José Caballero"
Dpto. Tecnología
Avda. Nuevo Colombino S/N, 21007 Huelva
diego@dartsystems.es

José Manuel Martín Ramos
Universidad de Huelva
Dpto. Tecnologías de la Información
E.P.S. "La Rábida"
Ctra. Huelva-La Rabida S/N, 21047 Huelva
jmmartin@dti.uhu.es

Resumen: En este artículo se presenta una propuesta de sistema de detección de plagio externo monolingüe basada en una modificación del concepto de n-grama ("n-grama contextual"), un nuevo motor de búsqueda basado en dicho concepto, y una nueva estrategia de determinación del plagio y sus límites ("monotonía referencial"). Los resultados de evaluación obtenidos son comparables a los del primer clasificado en la PAN'09, aunque obtenidos con un muy inferior coste computacional (tiempo de ejecución entre 30 y 45 minutos en un PC portátil sin uso de programación concurrente), lo que lo convierte en una muy interesante alternativa a explotar.
Palabras clave: detección de plagio, n-grama, n-grama contextual, Monotonía Referencial, Recuperación de Información.

Abstract: In this paper a new approach is shown for a monolingual extrinsic plagiarism detection system based on a modification of the "n-gram" concept (named "contextual n-gram"), a new high performance Information Retrieval engine based on this new concept, and a new strategy ("referential monotonity") for plagiarism detection and its limits. The assessment results can be compared with those results carried out by the winner team in PAN'09, but these are achieved with very low computational cost (results available between 30 and 45 minutes on a single laptop machine and without using concurrent programming) compared with the other existing works. Because of that, it is a very interesting proposal to exploit.
Keywords: plagiarism detection, n-gram, contextual n-gram, Referential Monotonity, Information Retrieval

1 Introducción

Con la aparición de Internet y la gran cantidad de documentos que ofrece a la disposición de sus usuarios, se hace patente un crecimiento del número del plagio de documentos, a su vez complicado de detectar por dicho extenso número de probables fuentes de plagio existentes.

Esto hace evidente la necesidad de medios automatizados que faciliten la detección de los posibles plagios para ser finalmente verificados por un especialista humano.

En este artículo se presenta un sistema de detección de plagio externo monolingüe basado en tres innovadoras propuestas:

- una modificación del concepto de n-grama, el ***n-grama contextual***, que reúne dos características: describe el contexto de la sentencia en la que se encuentra y es una huella o firma altamente discriminativa de la propia sentencia o del documento frente al resto en una extensa colección.
- un **nuevo sistema de Recuperación de Información** (RI) de altísima precisión específico para este fin, basado en el anterior concepto.

ISSN 1135-5948

- una **nueva estrategia** de determinación del plagio y sus límites, denominada ***Monotonía Referencial (MR)***.

Los resultados de evaluación obtenidos son comparables a los de los primeros clasificados en la PAN'09, obteniendo en general mejores resultados, con un coste computacional muy inferior (en torno a 30~45 minutos de tiempo de ejecución en un PC portátil sin uso de programación concurrente), lo que lo convierte en una muy interesante alternativa a explotar.

2 Estado del Arte

Puesto que este artículo no pretende entrar en detalles sobre los distintos enfoques y estudios existentes para el análisis de plagio y reuso de texto, referimos al lector a los trabajos de (Barrón 2008) y (Cough 2003).

2.1 PAN'09

PAN'09 [1] es una competición que aparece como actividad complementaria en la SePLN'09 [2] para la Detección de Plagio, constituyendo la más reciente muestra de estado del arte sobre la materia.

Dicha competición trataba la resolución de dos problemas diferentes: Detección Intrínseca de Plagio[1] y Detección Externa[2]. La propuesta de este artículo pertenece a este último tipo.

2.1.1 Corpus de desarrollo para PAN'09

Para llevar a cabo dicha competición, se ofreció a los interesados un excelente e imprescindible recurso con el que llevar a cabo el desarrollo de los sistemas de detección de plagio, consistente en un corpus de texto sin formato con 7214 documentos fuente (1GByte, 90% en inglés y 10% en alemán o español), independientes entre sí, y 7214 de documentos sospechosos (1 GByte solo en inglés), de los cuales, el 50% había sufrido plagio artificial a partir de los documentos fuente mediante un programa desarrollado a tal efecto llamado *"Plagiador Aleatorio"*. Dicho programa aplica cuatro tipos de plagio, en función del nivel de ofuscación y de la existencia o no de traducción.

[1] Identificar el plagio por cambios de estilo en la escritura, sin necesitar probables fuentes de referencia

[2] Identificar secciones plagiadas tanto en el documento sospechoso como en las probables fuentes, identificadas en una colección de referencia.

La colección de documentos sospechosos, se acompaña de información sobre los plagios realizados en XML.

Para la fase de competición, se proporcionó otro corpus de características similares, sin información XML de solución.

El sistema de cada equipo competidor, debía generar su resultado en XML y enviarlo al comité organizador, para evaluarlo frente al correcto, no revelado hasta el final de la competición.

2.2 PAN Plagiarism Corpus 2009

Tras la PAN'09, se liberó el nuevo corpus ***PAN-PC-09***[3] [3].Este es un nuevo recurso a gran escala que la Universidad de Weimar y la Universidad de Valencia, ponen a nuestra disposición para la evaluación controlada de algoritmos de detección de plagio .

Contiene 41.223 documentos de texto plano (obtenidos a partir de 22.874 libros de *Project Gutenberg* [4]) con 94.202 casos de plagio artificial insertados mediante el "*Plagiador Aleatorio*", y está preparado para evaluar tanto análisis externo (70%) como intrínseco (30%).

3 Proceso genérico para la detección de plagio externo

En (Potthast et al. 2009), se esboza el proceso genérico para la detección de plagio externo. Nuestro sistema se ajusta bastante al mismo.

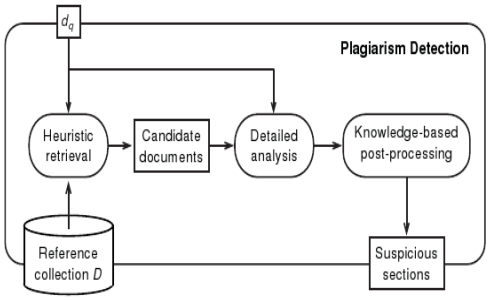


Figura 1: Proceso genérico de análisis de plagio externo (Potthast et al. 2009)

En la figura 1 se observa como para un documento sospechoso d_q, a través de una heurística de RI, se seleccionan una serie de documentos candidatos a ser su fuente de plagio. Mediante un proceso de análisis detallado de los pares (fuente y sospechoso) y

[3] http://pan.webis.de

un post-procesamiento basado en el conocimiento se obtienen las secciones, en ambos documentos, sospechosas de haber sufrido plagio.

4 "N-grama contextual" como base para el diseño de un sistema de R.I. específico.

El sistema que se presenta en este artículo basa el análisis y detección de plagio en la comparación de n-gramas (Barron y Rosso 2009a), usando un tratamiento especial para la construcción de los mismos.

La denominación "n-grama contextual" hace referencia a la característica de dichos n-gramas para describir la esencia del contexto en un muy reducido grupo de palabras.

La simple extracción de tokens para formar n-gramas hacen el análisis de plagio muy vulnerable a la ofuscación.

Para conseguir que el n-grama contenga la mejor definición de la esencia del contexto y sea especialmente útil para localizar posibles plagios, ofuscados o no, se llevan a cabo seis pasos para el modelado de los documentos en el proceso de análisis de plagio:

1. La **conversión a minúsculas** en la tokenización es una práctica común que obviamente se incluye.
2. La **eliminación de las palabras vacías** (*stopwords*), hacen que el n-grama contenga una información mucho más definitoria del contexto de la sentencia analizada con un mismo número de palabras. Si además tenemos en cuenta que los plagiadores suelen emplear eliminación y sustitución de palabras, los autores consideran que las *stopwords son* precisamente las más fáciles de eliminar y sustituir sin afectar al sentido/contexto original de la frase, por lo que no tiene mucho sentido conservarlas.
3. **Eliminación de tokens de un solo carácter**, ya que suelen corresponder a enumeraciones. Su elevada frecuencia de aparición hace considerarlos poco significativos. Este paso insensibiliza el sistema ante el posible cambio de orden en listas de sentencias.
4. **La reducción a la raíz** (*stemming)* [5] contribuye a la mejora de cobertura en la detección de plagios en los casos de sustitución de palabras por sus derivadas.
5. **Ordenación alfabética interna de los tokens del n-grama**, procesándose como representante canónico del conjunto de sus posibles permutaciones. Este paso anula el efecto de un posible cambio de **orden de las palabras** al reescribir la frase (como el cambio de activa a pasiva, etc.) o al traducir de un idioma a otro (plagio translingüe) (Potthast et al. 2010). Es patente que conservar el orden en la construcción de n-gramas, puede conducir a una pérdida importante de cobertura cuando el plagio tiene algún tipo de ofuscación.

6. El empleo de **solapamiento de n – 1 tokens** (en su orden natural) entre n-gramas contextuales consecutivos, permiten una mejor detección ante los intentos de ofuscación ya descritos.

Se podría pensar que todas estas técnicas que ayudan a mejorar la cobertura ante la posible ofuscación, podrían facilitar la aparición de falsos positivos, pero se demuestra experimentalmente que los pasos 2 y 3 compensan sobradamente la capacidad discriminatoria final del n-grama contextual.

Tras el estudio de los corpus del PAN'09, se ha comprobado que un alto porcentaje de este tipo de n-gramas tienen tal capacidad discriminatoria, que en la práctica constituyen una firma tanto del documento como de la sentencia que los contiene, especialmente para el caso de n-gramas de grado 3 y superiores (figuras 2-4).

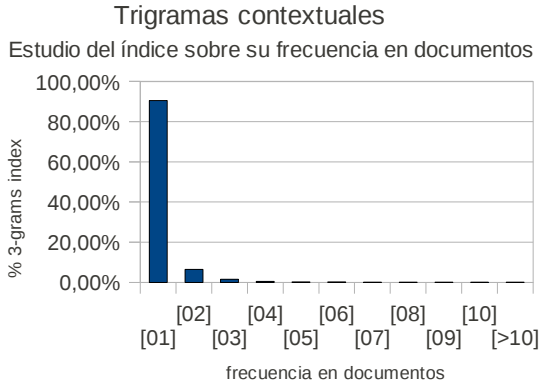
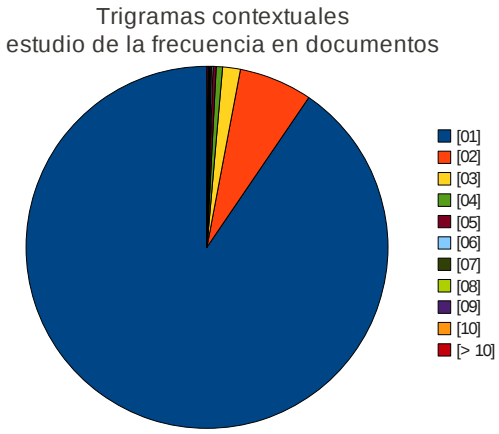
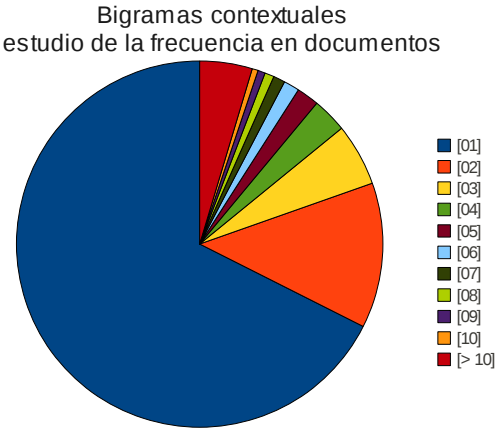


Figura 2: Corpus de desarrollo de PAN'09 – Índice de trigramas contextuales.





Figuras 3 y 4: Corpus de desarrollo de PAN'09 – Índices de n-gramas contextuales.

Esta capacidad discriminatoria se fortalece por la asociación de otros n-gramas contextuales cercanos, por lo que constituyen una base excelente para el desarrollo de un sistema de RI específico para la detección y localización de posibles plagios.

En el mismo estudio se observó que la inclusión o no de los números en la indexación de la misma colección afecta mínimamente al tamaño del índice, conservando en la práctica las mismas características y prestaciones.

Mediante el análisis del índice del corpus de desarrollo de la PAN'09, se calculó que la probabilidad de que un trigrama contextual de un nuevo documento no plagiado, se encuentre en un documento concreto de la colección, es del 0,0026%, mientras que la de que un trigrama contextual plagiado detectado apunte al documento fuente correcto usando sólo tres referencias por trigrama[4] es del 94,37%.

El mismo cálculo, basado en bigramas contextuales nos ofrece una probabilidad de repetirse en un documento concreto del 0,052% y la de apuntar al documento correcto en caso de coincidir y estar plagiado, es del 77,00% ó 77,76% con máximos de 5 y 9 referencias/bigrama respectivamente.

Estas características convierten al n-grama contextual en un medio excelente para calcular la similitud contextual entre pasajes o entre pasajes y documentos, capaz de identificar al más probable con un altísima precisión, seleccionando como consulta un pasaje sospechoso de 10 o más n-gramas.

El único inconveniente de esta capacidad discriminatoria, es el gran tamaño del Índice Invertido resultante, varias veces superior a la propia colección, especialmente si se incluye la información de su localización de los n-gramas (offset – length) y de su frecuencia de aparición dentro de cada documento, ya que es más probable que exista repetición de un n-grama dentro del mismo que en otro distinto.

Como uno de los objetivos de este trabajo es realizar un sistema rápido en una arquitectura de bajo coste (un PC con 4 GB de RAM), si se indexan n-gramas, resulta un lujo prohibitivo guardar dicha información para emplear el extendido modelo del *Espacio Vectorial*, calculando la similitud en base a la tf – idf.

Sin embargo, se ha comprobado que una ingeniosa aproximación basada solo en pesos proporcionales a la inversa de la *df*[5], limitando la concreción de referencias a documentos, es suficiente para obtener un sistema de RI de altísima precisión al que enviar como consulta pasajes de 'L' n-gramas contextuales por consulta (valores de L = 10 son más que suficientes, aunque los mejores resultados se obtuvieron con valores entre 25 y 45), por lo que para nuestro sistema basta con guardar la *df* y las primeras 3 referencias para trigramas y 5 ~ 9 referencias para bigramas.

Si se dispone de un sistema de RI específico con estas características, se puede aprovechar su precisión para estrechar el espacio de búsqueda (Barrón y Rosso, 2009b).

[4] Número máximo de referencias a documentos que guardará el sistema de RI. Por encima del mismo se consideran "demasiado frecuentes" y se desprecia su identificación, aunque no la frecuencia.

[5] Frecuencia de aparición de un término para la colección en base al nº de documentos.

Es bien conocido que el plagiador suele nutrirse de varias fuentes de plagio, por lo que es conveniente que el sistema devuelva un grupo de documentos candidatos (o pasajes de ellos) suficientemente amplio para abarcar en lo razonable dichas fuentes, analizarlas en profundidad y elegir las mejores.

Disponiendo de un sistema de RI como el propuesto en este artículo, es preferible un cambio de estrategia tras dividir el documento sospechoso en fragmentos pequeños: identificar un único mejor candidato a ser su fuente de plagio, obteniendo una lista de candidatos específicos (uno para cada fragmento) previo a la determinación de la posible zona de plagio.

5 *Monotonía Referencial*

En el análisis, es altamente probable que un un gran número de fragmentos sospechosos tengan asociado un posible documento fuente. Analizarlos todos, además de requerir un gran tiempo, provoca el aumento de falsos positivos.

Para evitarlo se recurre a una nueva estrategia, denominada Monotonía Referencial (MR), consistente en desechar aquellos fragmentos que aparecen aislados, sin repetirse al menos un determinado número de veces (umbral de monotonía). La MR consigue capturar los fragmentos que son suficientemente extensos como para indicar que existe una alta correspondencia entre los documentos analizados.

El la figura 5, se detalla el proceso de detección y reducción del espacio de búsqueda por el principio de MR: los splits[6] marcados en gris oscuro, son los que dan la detección directa (repetición de la referencia al documento 91 en 5 splits consecutivos) por igualar o superar el umbral de MR (en el ejemplo, 4 repeticiones).

73	-1	6	49	11	-1	31	91	91	91	91	91	6	92	5	7	98	91	57	-1	-1	-1	61

Figura 5: una única fuente candidata a estar plagiada por split, base de la estrategia de MR.

El sistema mostró un excelente comportamiento basándose solo en esta estrategia, con una longitud de split de 25 n-gramas contextuales y un umbral de MR de 4 repeticiones. Esta estrategia tiene por contra, el inconveniente de no detectar fragmentos plagiados que afecten a 3 o menos splits, (75 n-gramas contextuales que equivalen a unas 150 palabras en la práctica).

Otra mejora del algoritmo para detectar plagios de menor longitud, al menos para los plagios literales o escasamente ofuscados, es la reducción del umbral de monotonía cuando el índice de similitud supera el umbral de 0,75 para trigramas o 0,50 para bigramas. Valores inferiores de estos umbrales no compensan por la aparición de falsos positivos en cuanto a encabezamientos, frases comunes, etc.

6 *Reducción del tiempo de cómputo.*

El tiempo de computo ha sido desde el inicio un objetivo fundamental en el desarrollo del sistema de análisis. En experiencias iniciales basadas en n-gramas, usando un buscador genérico (*Lucene*) e implementando el sistema en Java, se estimaron tiempos de computo cercanos a 27 días en un PC para obtener los primeros resultados de análisis de la colección completa.

La minimización de este tiempo de computo se ha llevado a cabo utilizando diferentes estrategias software como el empleo de un Árbol Binario de Búsqueda (ABB) como medio más eficiente para la ordenación de los n-gramas conforme eran extraídos de los documentos fuente.

El enorme gasto de memoria necesario para el ABB hizo necesaria la aplicación de otra estrategia software, consistente en el uso de un array (obtenido de la mezcla de recorridos *inorden* de ABB parciales) para la construcción del Índice Invertido del sistema de RI.

La utilización del array en vez del ABB se debe a que necesita menos memoria, dando el mismo orden de eficiencia al emplear un algoritmo de búsqueda dicotómica o binaria.

Se ha comprobado que con estas estrategias, se necesita algo menos de 3 GB de índice (de trigramas contextuales) para 1GB de probables fuentes en texto plano.

Las mejoras de tiempo de computo obtenidas utilizando las estrategias de programación, los n-gramas contextuales, el nuevo motor de RI y la poda por MR han conseguido tiempos inferiores a una hora. En concreto se consigue realizar el análisis del corpus completo de desarrollo (1GB + 1GB txt) en 27 minutos utilizando trigramas y en 21 minutos para bigramas (más el tiempo de indexación del corpus de documentos fuente, de

[6]Divisiones del documento que contienen un número fijo de n-gramas contextuales consecutivos.

15 y 8 minutos para trigramas y bigramas respectivamente).

7 Sistema de detección de plagio

Los pasos generales que realiza el sistema para el proceso de análisis externo son:

1. Clasificación de los documentos de la colección en base al idioma.
2. Construcción del Índice monolingüe, volcado a disco y carga del mismo en memoria.
3. División de los documentos sospechosos en splits con un número determinado de n-gramas contextuales.
4. Recuperación de un único documento fuente para cada split mediante el nuevo sistema de RI.
5. Determinación de la existencia de plagio basada en la monotonía de aparición de referencias a la misma fuente para splits consecutivos (MR).
6. Determinación de los límites de plagio por separado para el documento sospechoso, mediante una doble búsqueda (desde el inicio y el final del fragmento detectado) de los n-gramas comunes fronterizos.
7. Utilización de la zona sospechosa detectada para buscar en el documento fuente los límites de la zona con mejor correspondencia. En este paso se obtiene mejor precisión y cobertura que en la zona de plagio detectada previamente. Por ello se realiza un posterior refinamiento de los límites del fragmento sospechoso.
8. Grabación de los resultados en XML.
9. Evaluación sobre el conjunto de entrenamiento.

8 Ensayos y resultados obtenidos

Durante la gestación del sistema, se realizaron más de 150 pruebas con el corpus de desarrollo de la PAN'09, empleando n-gramas contextuales de grados 2 y 3, y ajustando distintos parámetros hasta llegar a los considerados como óptimos.

Las mejores prestaciones[7] se obtuvieron con trigramas, aunque el análisis fue algo más lento y necesitó más recursos que con bigramas.

En la tabla 1, se muestran los mejores resultados obtenidos para trigramas sobre el corpus de desarrollo PAN'09 al completo.

Trigramas (long. split: 25, umbral MR: 4)	
Precision	0,7915
Recall	0,6370
F-measure	0,7059
Granularity	1,0136
Overall	0,6991

Tabla 1: Prestación monolingüe - trigramas corpus desarrollo PAN'09

Analizando la misma colección en 10 folds[8] mediante trigramas, se obtuvo una desviación típica del 1,53% y máxima del 3,99%.

El mismo análisis pero con bigramas, obtuvo una desviación típica del 2,70% y máxima del 4,48%. La tabla 2 muestra las prestaciones para el corpus de desarrollo.

Bigramas (long. split: 30, umbral MR: 4)	
Precision	0,7256
Recall	0,6013
F-measure	0,6576
Granularity	0,0155
Overall	0,6504

Tabla 2: Prestación monolingüe – bigramas corpus desarrollo PAN'09

Los resultados de las tablas 3 (trigramas) y 4 (bigramas) se obtuvieron analizando el corpus de competición con los ajustes obtenidos a partir del de desarrollo, teniendo en cuenta la penalización implícita por no analizar el plagio translingüe presente en el mismo, lo que permite comparar nuestro sistema con la muestra de estado del arte que supuso la PAN'09.

Trigramas (long. split: 25, umbral MR: 4)	
Precision	0,7989
Recall	0,6349
F-measure	0,7075
Granularity	1,0142
Overall	**0,7003**

Tabla 3: Prestación multilingüe – trigramas corpus competición PAN'09

Como se observa, las prestaciones empleando trigramas contextuales son incluso

[7] http://www.uni-weimar.de/medien/webis/research/workshopseries/pan-10/task1-plagiarism-detection.html#measures

[8] Subconjuntos de similar tamaño obtenidos de dividir el corpus de sospechosos.

algo superiores a las del equipo ganador (Grotzea et al 2009), además de obtenerse en una máquina de prestaciones muy inferiores y en un tiempo de computo mucho menor.

Bigramas (long. split: 30, umbral MR: 4)	
Precision	0,7375
Recall	0,5966
F-measure	0,6596
Granularity	1,0181
Overall	**0,6511**

Tabla 4: Prestación multilingüe – bigramas corpus competición PAN'09

En las siguientes figuras se muestran los resultados obtenidos por los participantes en PAN'09, y se han insertado los obtenidos por la propuesta presentada, empleando trigramas (T) y bigramas (B) contextuales.

Las figuras 6 y 7 muestran las gráficas comparativas del sistema presentado frente a los mejor clasificados en la PAN'09.

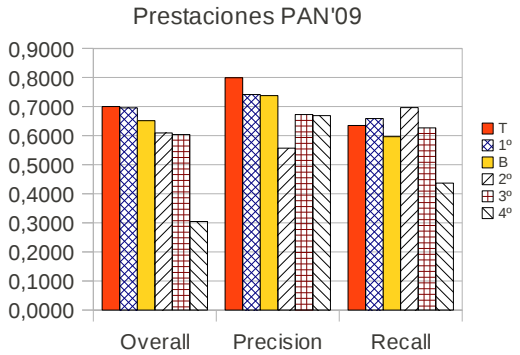


Figura 6: Comparativa con los 4 primeros clasificados PAN'09

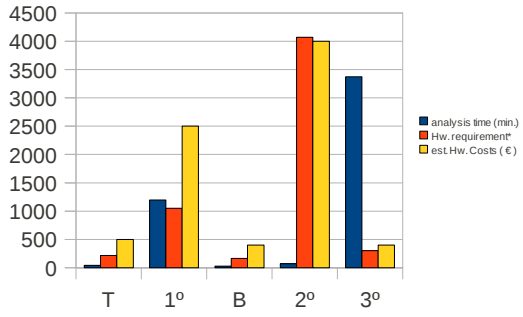


Figura 7: comparativa: Tiempos de análisis - Requisitos Hardware – Costes (a)

En las figuras 8~13 se muestran las gráficas correspondientes a las comparativas del sistema presentado frente a los de los participantes en la PAN'09. La comparativa de tiempos y requisitos hardware frente a los tres primeros clasificados (figuras 7 y 13), se estimó (Rodríguez-Torrejón D.A. 2009) en base a la información disponible en los artículos de la PAN'09.

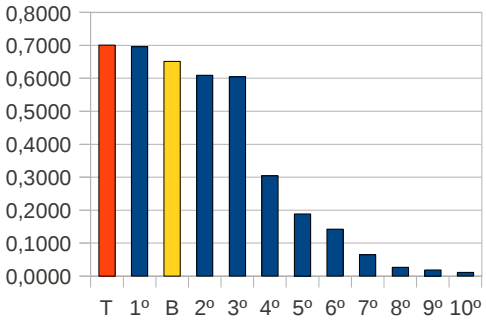


Figura 8: Puntuación General (Overall)

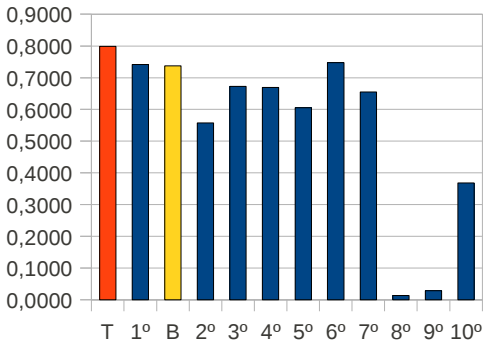


Figura 9: Comparativa de Precisión

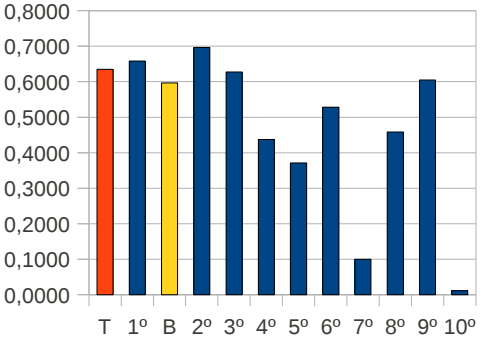


Figura 10: Comparativa de *Recall*

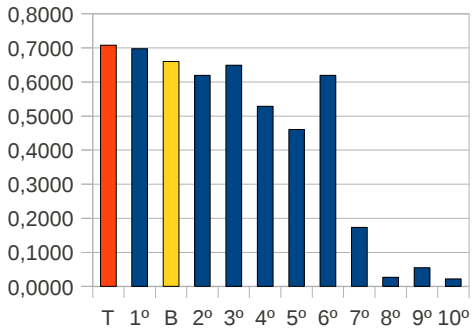


Figura 11: Comparativa de F-measure

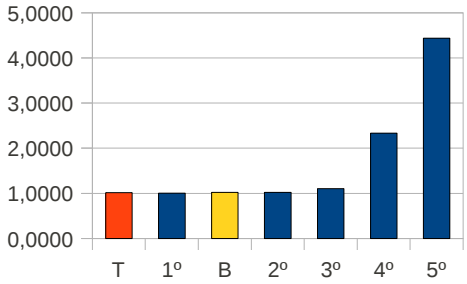


Figura 12: Comp. de Granularidad

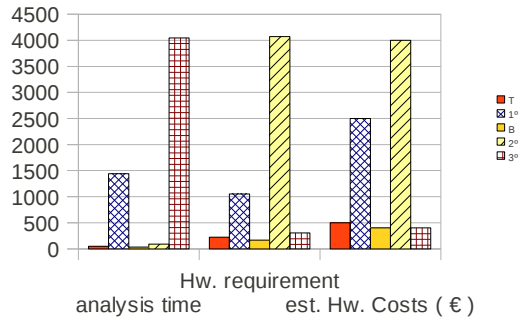


Figura 13: Comparativa escalada: Tiempos de análisis - Requisitos Hardware – Costes

9 *Conclusiones y futuros trabajos*

Se considera que los resultados obtenidos con el sistema presentado en este articulo son muy buenos y prometedores si los comparamos con los sistemas presentados en PAN'09 además de poseer una notable mejora respecto a las necesidades hardware y los tiempos de computo necesarios.

Existen muchas más estrategias para la mejora del sistema presentado. Una posible mejora, que no ha sido utilizada para obtener los resultados presentados, es la de realizar un análisis mas exhaustivo de la valiosa información (obtenida del algoritmo de MR) que supone conocer las fuentes de plagio de los tramos medianos y grandes. En la figura 15 se puede observar que un segundo recorrido mediante *feedback* (zona de la derecha) por la lista de referencias, se pueden contemplar los splits que apunten a documentos detectados previamente (zona de la izquierda), pues muchos plagiadores suelen recurrir a la mismas obras para extraer varios fragmentos.

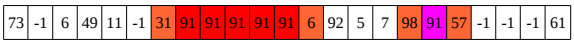


Figura 15: Mejora de MR mediante *feedback*

Existen muchas mas estrategias que pueden ser utilizadas para complementar y mejorar los resultados obtenidos en este artículo, entre las cuales podemos citar:

- Sistemas multilíngües.
- Sistemas concurrentes y multinúcleo.
- Sistemas utilizando supercomputación.

Además quedan por explorar otros usos del n-grama contextual en el PLN, tales como búsqueda orientada a respuestas, clustering, clasificación, etc.

10 *Agradecimientos*

Los autores agradecen la inestimable ayuda de los investigadores D. Alberto Barrón y Dr. D. Paolo Rosso (U.P. Valencia), por su desinteresada colaboración para la validación del sistema de evaluación de prestaciones del sistema presentado.

También desean agradecer a todo el equipo de desarrollo de los corpus PAN'09 y PAN-PC-09, por facilitar estos magníficos e imprescindibles recursos de desarrollo, sin los que no habrían podido llegar a la obtención de estos resultados, y al resto de organizadores y participantes de dicha competición, pues su trabajo ha supuesto una fuente de motivación para continuar abordando el problema.

Webgrafía

[1]*PAN'09* (Workshop sobre el análisis plagio enmarcado en la SPLN '09) http://www.uni-weimar.de/medien/webis/research/workshopseries/pan-09/index.html

[2]*sePLN'09* (Congreso de la SEPLN en 2009) http://ixa2.si.ehu.es/sepln2009

[3]Webis at Bauhaus-Universität Weimar and NLEL at Universidad Politécnica de Valencia. *PAN Plagiarism Corpus*

PAN-PC-09 http://pan.webis.de 2009. M. Potthast, A. Eiselt, B. Stein, A. Barrón Cedeño, and P. Rosso (editors).

[4]*Proyecto Gutenberg* http://www.gutenberg.org

[5]*Martin Porter Stemming Algorithm* http://tartarus.org/~martin/PorterStemmer/index.html

Bibliografía

Barrón-Cedeño A. 2008. Detección automática de plagio en texto. Tesis de Máster - Universidad de Valencia.

Clough P. 2003. Measuring Text Reuse. PhD Thesis - University of Sheffield.

Potthast M., Stein A., Eiselt A., Barrón-Cedeño A., Rosso P. 2009. Overview of the 1st International Competition on Plagiarism Detection. En:

Stein B., Rosso P., Stamatatos E., Koppel M., and Agirre E., editors. *SEPLN 2009 Workshop on Uncovering Plagiarism, Authorship and Social Software Misuse (PAN 09), pp. 1-9, Donostia-San Sebastian, Spain, September 2009*. CEUR-WS.org. ISSN 163-0073.

Barrón-Cedeño A. y Rosso P. 2009a. On Automatic Plagiarism Detection based on n-grams Comparison. *Proc. European Conference on Information Retrieval, ECIR-2009,Springer-Verlag, LNCS* (5478) páginas 696-700.

Potthast M., Barrón-Cedeño A., Stein B., Rosso P. 2010 (en prensa). Cross-Language Plagiarism Detection. Languages Resources and Evaluation (Special Issue on Plagiarism and Authorship Analysis). DOI: 10.1007/s10579-009-9114-z

Barrón-Cedeño, A. y Rosso P. 2009b. On the Relevance of Search Space Reduction in Automatic Plagiarism Detection. *Procesamiento del Lenguaje Natural*, 43:141-149.

Grozea, C., Gehl C. y Popescu M. N. 2009. ENCOPLOT pairwise sequence matching linear time plagiarism detection (PAN'09 papers).

Rodríguez-Torrejón D. 2009. Detección de plagio en documentos. Propueta de sistema externo monolingüe de altas prestaciones basada en n-gramas. Tesina fin de Máster – Universidad de Huelva