Representación basada en grafos para la identificación de autoría en textos para el idioma español

Nahun Loya[1], Iván Olmos[1], David Pinto[1] y Jesús González[2]

[1]Benemérita Universidad Autonóma de Puebla,
Facultad de ciencias de la computación, México

[2]Instituto de Astrofísica, Óptica y Electronica,
Departamento de ciencias de la computación, México
{nahun.loya,iolmos,dpinto}@cs.buap.mx, jagonzalez@inaoep.mx

Resumen El proceso de identificación de autoría consiste en determinar al autor que ha escrito algún tipo de documento. Para realizarse es necesario conocer las características más adecuadas que permiten identificar los rasgos de escritura. Este trabajo propone una metodología basada en representaciones de textos a través de grafos etiquetados. Los grafos son utilizados como elemento esencial para la búsqueda de patrones o características. Se trabaja con un conjunto de datos que corresponde al conjunto de documentos (libros) pertenecientes a seis autores cada autor con cinco documentos diferentes. Se realiza una fase de preprocesamiento consistente en la eliminación de palabras cerradas y signos de puntuación, posteriormente los documentos son separados en frases. Dichas frases son representadas a través de grafos etiquetados, haciendo uso de dos representaciones propuestas. El objetivo es extraer características importantes. Dichas características son usadas en el proceso de clasificación. Los resultados muestran una mejora tomando como métrica los niveles de precisión y en comparativa con otros métodos como la bolsa de palabras o las características estilométricas.

Palabras clave: atribución de autoría, representaciones de texto en grafos, autoría en español, clasificación.

1. Introducción

La atribución de autoría suele verse como un tema de la lingüística computacional y tiene como objetivo identificar al autor o escritor original de un texto. Para lograrlo es necesario buscar características o perfiles que identifiquen plenamente al autor. Esta no es una tarea trivial debido a que los estilos de escritura suelen ser similares, inclusive los textos suelen ser escritos por diversos autores lo que dificulta aún más el proceso de búsqueda de patrones.

Actualmente se han estudiado diferentes técnicas para solventar el problema, las cuales van desde aplicaciones matemáticas que intentan medir el grado de frecuencia haciendo uso de métricas tradicionales como la varianza y la desviación estándar y modelos de aprendizaje automático.

En lo que compete al ámbito lingüístico computacional se han buscado características que van desde examinar las palabras en el texto, métricas como el tamaño de la letra, el número de vocales o consonantes utilizadas e inclusive se han realizado análisis de errores sintácticos y ortográficos.

El interés en esta área actualmente radica en que existen diversos documentos que no tienen asociado un autor o son considerados anónimos. Por tanto existe una gran necesidad de desarrollar modelos automáticos y confiables para la extracción de características que coadyuven a la solución óptima del problema.

Ahora que se ha establecido la problemática, en este sentido se presenta un modelo de representación de texto mediante grafos. Los grafos son una manera natural de representar la información, dado que se puede establecer conceptos a través de los nodos y relaciones a través de las aristas. Los documentos se pueden dividir en párrafos y estos en enunciados. En estos existen relaciones entre las palabras tal y como los nodos de los grafos son ligados a través de aristas. Hablamos de grafos etiquetados en el sentido de hacer una extensión de la noción típica de grafo (conjunto de vértices y aristas). Esta extensión permite incorporar relaciones habituales en el texto a través de los grafos.

El caso de estudio que se plantea considera 6 autores literarios, cada uno con cinco documentos (libros) diferentes pertenecientes al género de aventura. Se pretende mediante este conjunto de datos evaluar la calidad de la metodología la cual tiene como idea fundamental el uso de los grafos como herramienta de representación.

El resto del artículo se encuentra conformado de la siguiente manera. En la sección 2 se hace una revisión de los trabajos relacionados considerando dos aspectos: la atribución de autoría y el uso de grafos en el area de lingüística computacional. En la sección 3 se presenta una notación que es empleada para comprender el tipo de estructura de grafos propuesta en este trabajo. En la sección 4 se presentan dos propuestas que son usadas para la extracción automática de características, además de establecer una justificación del uso de las representaciones. En la sección 5 se detalla la metodología usada en este trabajo así como se presenta el caso de estudio. En la sección 6 se presenta un análisis de resultados, así como una comparativa entre los resultados obtenidos con los algoritmos de clasificación y un contraste entre los modelos de grafos y otros similares. Finalmente en la sección 7 se presentan las conclusiones y el trabajo a futuro.

2. Trabajo previo

El trabajo de atribución de autoría se plantea como un proceso típico de clasificación donde es necesario extraer características y las clases son los diferentes autores. Los primeros modelos para solventar el problema plantean el uso de bolsa de palabras como conjunto de características; en este caso se considera cada palabra del vocabulario como un atributo [6,13]. La desventaja radica en que estos modelos no muestran las relaciones entre el texto, enunciados o palabras. Otros enfoques similares tratan de extraer características a partir de las

palabras (total de palabras, tamaño de las palabras, número de vocales, número de consonantes, signos de puntuación, etc.), que al final son representadas como vectores de soporte [10]. Enfoques minuciosos son los relacionados con el análisis de las partes de la oración (adverbios, verbos, adjetivos, etc.) [8]. Estos incorporan información relacionada con la estructura del lenguaje empleado como lo son elementos léxicos, sintácticos [7]. Un método frecuente es el uso de n-gramas de palabras o de caracteres, donde se busca encontrar la estructura de la escritura [9]. Un trabajo sobresaliente es el mostrado en [5], donde se hace la evaluación de 39 modelos o procedimientos para caracterizar textos, dicha evaluación es realizada sobre un corpus formado por textos de una serie de columnistas del diario "Telegraphe" de Londres. Los resultados obtenidos muestran el porcentaje de confiabilidad de distintos métodos (palabras y marcas de puntuación, bigramas y trigramas de palabras, etc.) en comparación con la cantidad de autores (clases). Como resultado de este estudio se observa que los métodos obtienen una confiabilidad de hasta el 88 % cuando se consideran dos autores y baja radicalmente hasta 34 % cuando se trabaja con más de 10 autores.

En lo que concierne al tema de uso de grafos en el procesamiento del lenguaje natural, sin lugar a dudas son una buena forma de representación de dominios textuales. En la literatura distintas investigaciones relacionadas con la minería textual proponen el uso de los grafos. Desde la década de los 80's los grafos han sido usados para representar relaciones y conceptos. Basta con citar trabajos como los de [12], pionero en proponer los grafos conceptuales como un modelo para representar conocimiento. Estos grafos son concebidos de una forma psicológica, lingüística.

En particular en lo que concierne a la aplicación de determinación de autoría, los grafos también han sido utilizados con el objetivo de extraer características determinantes para la detección del autor. Ejemplo de este tipo de aplicaciones se muestran en [2], quien a partir de la representación de grafos de escritura logra obtener características importantes para la detección de autoría.

Otro estudio es el desarrollado en [11], donde se hace uso de parser de dependencias de Stanford para construir árboles que representan las relaciones sintácticas entre palabras denominadas "sn-gramas". Se muestra que la diferencia entre los n-gramas tradicionales y los sn-gramas radica en la forma en como son considerados los vecinos. Los "sn-gramas" son aplicados para extraer características del texto para atribución de autoría de tres autores logrando resultados favorables de hasta el 100 %.

Como se puede observar, en este análisis realizado existen diferentes propuestas para representar dominios textuales, con ventajas y desventajas. En este sentido, este trabajo de investigación explora una alternativa que permita abordar la representación de información en textos a través de grafos tomando como caso de estudio la atribución de autoría.

En la siguiente sección se presentan las nociones teóricas para la representación de grafos etiquetados. Se realiza una extensión de la notación tradicional de grafos con el objetivo de representar grafos etiquetados.

3. Extracción de características y representación basada en grafos

Tomando como base la noción de grafo, los grafos son representados por una dupla de la forma $G = (V, E)$, donde V es el conjunto de vértices del grafo y $E \subseteq V \times V$ representa la asociación entre vértices mediante arcos. Sin embargo, esta notación no es útil para el caso de representación de conocimiento, ya que no solamente es de nuestro interés guardar la estructura del grafo sino además el contenido de la información que se quiera mapear (etiquetas a los vértices y a los arcos), por ello introducimos la siguiente notación.

Definición 1 *Grafo. Es una sextupla donde:* $G = (V, E, L_V, L_E, \alpha, \beta)$

- $V = \{v_i | i = 1, ..., n\}$ es un conjunto finito de vértices, $V \neq \emptyset$, y $n = \#vertices$ en el grafo.
- $E \subseteq V \times V$ es el conjunto finito de aristas, $E = \{e = \{v_i, v_j\} | v_i, v_j \in V, 1 \leq i, j \leq n\}$.
- L_V, es un conjunto de etiquetas para los vértices.
- L_E, es un conjunto de etiquetas para las aristas.
- $\alpha : V \rightarrow L_V$, es una función que asigna las etiquetas a los vértices.
- $\beta : E \rightarrow L_E$, es una función que asigna etiquetas a las aristas.

En esta representación se asume que no existe una dirección entre el vértice origen (v_o) y el vértice destino (v_d) que unen cada arco. Si se quisiera representar un grafo dirigido se puede utilizar la notación para un arco $\{e = (v_o, v_d)\}$ (en lugar de utilizar llaves), indicando que los paréntesis establecen un orden.

Una vez introducida la notación en la subsecuente sección se presentan dos propuestas basadas en grafos etiquetados para la representación de texto.

4. Representaciones propuestas

Las siguientes representaciones toman un texto que es divido por frases y para cada frase se construye un grafo.

4.1. Representación A

Se basa en establecer las relaciones entre palabras de cada enunciado. Para cada palabra le es asignado un nodo del grafo, este es etiquetado con la misma palabra. Con esta representación se busca mantener la secuencia de los términos en un enunciado. En esta representación las aristas son etiquetadas con la palabra genérica "siguiente"de acuerdo al grado de aparición de cada palabra. La representación no considera una topología base, más bien se adecúa a la estructura de la frase. Con ello se espera encontrar conjuntos de palabras importantes en el texto. En un comparativo con el modelo de n-gramas de palabras no se define un n fijo, más bien es la herramienta de minado la que establece y adapta ese n. Para ilustrar la representación contémplese el siguiente estracto de la

novela de Julio Verne "Viaje al centro de la tierra" "*...me quedé sólo, se me ocurrió la idea de írselo a contar a mi tío, si me quedé sólo ¿Y si mi tío volvía y me llamaba?...*". La representación basándose en la representación A es la que se observa en la figura 1.

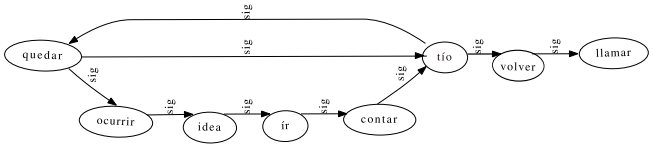

Figura 1. Representación A. Secuencia y orden de las palabras.

4.2. Representación B

La representación B es realizada en base a un etiquetador de dependencias de tipo. Un etiquetador de dependencias de tipo provee una descripción simple de las relaciones en una oración, mostrando las relaciones entre pares de palabras mediante tripletas de la forma *relación(palabra1, palabra2)*. Para realizar este tipo de representación se considera una topología base, el grafo en forma de estrella. La idea principal de esta representación consiste en que los lemas de las palabras son representados como nodos del grafo concatenandoles la clase gramatical, por otra parte las relaciones de dependencia de tipo son las etiquetas de las aristas. Esta representación anula la relación "siguiente" mostrada en la representación A y la substituye por la de dependencias de tipo. En un comparativo con la representación mostrada en [11], en este trabajo usamos el etiquetador de dependencias de tipo para encontrar las relaciones entre pares de palabras tomando considerando una topología base. Esto es con el objetivo de encontrar patrones de escritura con mayor cantidad de estructura y soporte que determinen las características del escritor. Refiérase nuevamente al enunciado anterior, un grafo etiquetado bajo la representación B se observa en la Figura 2.

Una vez que se ha presentado las representaciones se establece la metodología seguida a través de este trabajo en la siguiente sección.

5. Metodología y herramientas utilizadas

Antes de pasar a describir la metodología empleada se describe el conjunto de datos o corpora.

5.1. Conjunto de datos

El conjunto de datos esta conformado por los escritos literarios en español de los autores: C.S. Lewis, Darren Shan, J.K. Rowling, Justin Somper, Julio Verne

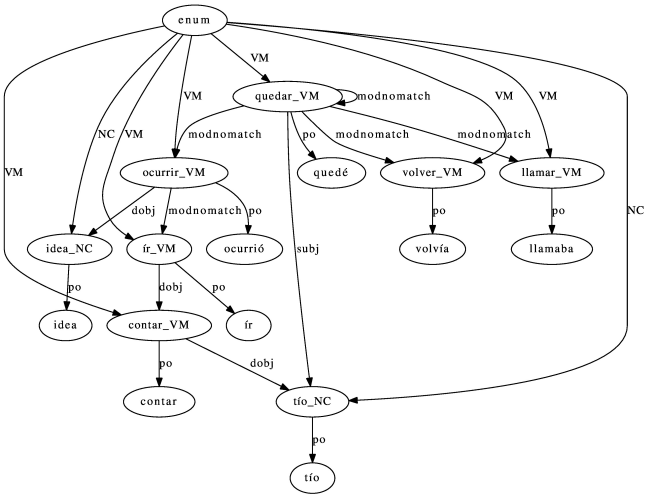

Figura 2. Representación B. Estructura de estrella utilizando un etiquetador de dependencias de tipo.

Tabla 1. Características del corpora considerado.

Característica	C.S. Lewis	Darren Shan	J.K. Rowling	Justin Somper	Julio Verne	Rick Riordan
Num. libros por autor	5	5	5	5	5	5
Tamaño del Vocabulario	37994	29865	88048	33802	72622	44879
Num. de frases	9191	8100	55934	54255	13454	20712
Promedio de palabras por libro	65472	39917	223386	130649	108874	54164

y Rick Riordan, pertenecientes a la literatura de aventura. Un panorama general de las características del corpus se muestra en la tabla 1.

Ahora se describe la metodología empleada que va desde los documentos originales hasta el planteamiento de los modelos de clasificación.

5.2. Metodología

El conjunto de datos es dividido por autor para su tratamiento. La serie de pasos considerado es el siguiente.

1. Cada corpus por autor es separado por enunciados. Entendiendo el concepto de enunciado como una serie de palabras que son separadas por un signo terminal, en este caso el punto.
2. Cada oración es limpiada de signos de puntuación adicionales.
3. Para cada oración se elimina el conjunto de palabras cerradas.
4. Cada oración es pasada por un etiquetador de partes de la oración para obtener el lema de cada palabra (Nota: para el caso de la Representación B se usa un etiquetador de dependencias de tipo.)

5. Se construye un grafo por oración.
6. Se agrupan el conjunto de grafos generados por autor.
7. Las características con mayor relevancia son extraídas del conjunto de grafos usando una herramienta de Minado en grafos.
8. Las características extraídas son colocadas como características de un clasificador supervisado usando Weka[1].

Como etiquetador de partes de la oración se usa Freeling[2] para el caso de la representación A. En el caso de la representación B se usa el mismo sistema (Freeling), sin embargo se utiliza el parser de dependencias (incluido en la herramienta). Para cada conjunto de oraciones se genera un conjunto de grafos basado en las representaciones propuestas. Este conjunto de datos es importante ya que es la entrada para la herramienta de minería en grafos. La salida de la herramienta de minado es la entrada como características para realizar una clasificación supervisada[3].

5.3. Herramientas utilizadas

Con el fin de extraer los patrones de escritura se utiliza la herramienta de minería de datos basada en grafos Subdue. Existen diferentes tipos de herramientas para obtener subestructuras comunes en un conjunto de datos (Subdue [3], gSpan [14], etc.), en este caso de grafos. En este trabajo se usa la herramienta Subdue debido a que su código es libre y ha sido probado en diferentes tareas: descubrimiento de patrones en telecomunicaciones [1], detección de anomalías [4], etc.

Subdue Es usada para descubrir conocimiento, encontrar estructuras y patrones relacionales a partir de la representación mediante grafos etiquetados muy útil en tareas donde se trabaja con dominios estructurados. Subdue usa tres métricas: MDL, la métrica "size" que reporta las mejores subestructuras basándose en el tamaño del grafo y la métrica "cover" que se basa en el número de repeticiones de las subestructuras en un conjunto de grafos.

Ahora que se han descrito todos los elementos necesarios realizados para la extracción de características se presentan los resultados obtenidos en términos del uso de diferentes clasificadores y con cada una de las representaciones propuestas.

6. Resultados obtenidos

La presente sección muestra los resultados de aplicar los métodos tradicionales tanto de características estilométricas (se consideraron: promedio de palabras entre oraciones, tamaño del vocabulario, tamaño de las palabras, número

[1] http://www.cs.waikato.ac.nz/ml/weka/
[2] http://nlp.lsi.upc.edu/freeling/
[3] Una descripción de las categorías gramaticales se muestra en la página, http://nlp.lsi.upc.edu/freeling/doc/tagsets/tagset-es.html/

de mayúsculas y minúsculas, número de signos de puntuación) como el modelo de bolsa de palabras ponderado por frecuencia como comparativo con el modelo propuesto de extracción de características basada en grafos. Las palabras obtenidas con la herramienta de minado son tomadas como características para el clasificador. Se usa la herramienta Weka, en la cual se prueban los conjuntos de características en diversos clasificadores supervisados, entre ellos: Random Forest, Máquinas de vectores de soporte (SVM), árboles de decisión C4.5 y el algoritmo de voto que incorpora la opinión de varios clasificadores. Para el caso particular de este último se consideran dos clasificadores: Random Forest y SVM; se eligen estos dos debido son estos los que aportan una mejor precisión en pruebas individuales. Los resultados están expresados por niveles de precisión. Cabe destacar que los resultados en la tabla 2 reflejan un análisis minucioso de la configuración de los clasificadores, es decir, se prueban diferentes configuraciones para cada uno y los resultados son reportados con las mejores. Se lleva a cabo una prueba de validación cruzada de 10 divisiones como conjunto de prueba. El corpus que se trabaja es el mostrado en la sección 5.1 y los resultados son los siguientes:

Tabla 2. Resultados obtenidos.

Clasificador	Bolsa de palabras (1000 atri.)	Características Estilométricas-6	Representación A	Representación B
Random Forest	0.71	0.65	0.78	**0.81**
SVM	0.75	0.65	**0.79**	0.80
C4.5	0.68	0.50	0.63	0.67
Multilayer Pe.	0.66	0.55	0.62	0.67
Voto	**0.73**	**0.69**	**0.79**	**0.81**

En la siguiente subsección se presenta un análisis meticuloso de los resultados obtenidos resaltando por un lado los niveles de precisión obtenidos por los clasificadores y por otra parte efectuando un comparativo entre los modelos de representación de características.

6.1. Análisis de resultados

Los resultados que figuran en la Tabla 2 muestran la confiabilidad de los clasificadores SVM y el algoritmo de árboles de decisión de Random Forest ya que estos obtienen la mayor precisión global. Enfatizando en las representaciones se muestra que el modelo basado en características estilométricas únicamente logra clasificar correctamente al 65 % de los casos cuando se usa una máquina de soporte vectorial y con otros clasificadores un nivel inferior. Los resultados de las características estilométricas se ven fácilmente superados por el modelo basado en bolsa de palabras cuyo máximo de elementos adecuadamente clasificados se observa en el algoritmo de Voto logrando un 69 % global. En lo que respecta a los resultados obtenidos usando la representación A estos muestran una mejora de

un 6 % en comparación con el mejor resultado obtenido por la bolsa de palabras. Se cree que esto es debido a que la representación A ayuda a plasmar la forma de escritura del autor mostrando las conexiones entre conjunto de palabras. Por otra parte en lo que concierne con la representación B los mejores resultados son los obtenidos con SVM y Random Forest, ya que bajo esta representación se están encontrando características que no sólo reflejan la secuencia de escritura del autor si no que incorporan características como las dependencias de tipo asociadas entre palabras incluyendo patrones léxicos y sintácticos. Esta última representación logra obtener hasta un 81 % en una prueba de validación cruzada.

En la siguiente sección se presentan las conclusiones de este trabajo, así como se realiza una serie de propuestas factibles para seguir explorando la riqueza de una representación basada en grafos.

7. Conclusiones y trabajo futuro

Se presentó una metodología que establece una aproximación para la solución del problema de determinación de autoría. En este trabajo se considera el conjunto de datos de seis autores cada uno con cinco documentos (en este caso libros). La idea fundamental consistió en realizar un proceso de descubrimiento de características fundamentado en la representación de textos basada en grafos. Se han establecido dos propuestas, por una parte la representación A conserva características de secuencia entre los enunciados y por otro lado una representación B la cual incorpora elementos léxicos, sintácticos y las relaciones (aristas) se establecen mediante el uso de las dependencias de tipo. Se realiza una clasificación con base en dos modelos tradicionales bolsa de palabras y características estilométricas de escritura, con el objetivo de realizar un comparativo con los modelos propuestos a través de grafos. Los resultados en términos de precisión muestran una mejora notable al hacer uso de las representaciones propuestas.

Sin lugar a dudas las representaciones basada en grafos fueron un elemento determinante en la búsqueda de características. A través de la minería de datos basada en grafos se pueden encontrar elementos que funjen como atributos para la clasificación. Se puede equiparar las representaciones con los modelos basados en n-gramas de palabras. Sin embargo, los grafos son capaces de proveer combinaciones de características que estén formadas no sólo por unigramas, bigramas o trigramas, sino más bien pueden proporcionar combinaciones notables de características. Además se enriquece la representación incorporando relaciones entre las palabras como es el caso de la representación B, a diferencia de los modelos típicos.

Estas representaciones aún pueden ser mejoradas estudiando las distintas topologías existentes. Además se pueden incorporar conceptos como la sinonimia, hiperonimia, aún no consideradas en estas representaciones. Se debe seguir indagando acerca de las distintas formas en que se pueden relacionar los nodos (palabras), ya que en una representación basada en grafos de esta forma se determina el grado de representatividad de escritura de determinado autor. Considérese que para este trabajo utiliza las representaciones basadas en grafos

como un método para extraer información del texto, por tanto se debe explorar la amplia gama de aplicaciones que las representaciones pueden tener, puesto que en este caso se observó únicamente para el problema de atribución de autoría.

Referencias

1. Baritchi, A., Cook, D.J., Holder, L.B.: Discovering structural patterns in telecommunications data (2000)
2. Castillo, E.: Determinación de características en el proceso de deteccion de autoría. FCC-BUAP (2012)
3. Cook, D.J., Holder, L.B.: Substructure discovery using minimum description length and background knowledge. J. Artif. Int. Res. 1(1), 231–255 (Feb 1994)
4. Eberle, W., Holder, L.: Anomaly detection in data represented as graphs. Intell. Data Anal. 11(6), 663–689 (2007), http://portal.acm.org/citation.cfm?id=1368024
5. Grieve, J.: Quantitative authorship attribution: An evaluation of techniques (2007)
6. Houvardas, J., Stamatatos, E.: N-gram feature selection for authorship identification. In: AIMSA. pp. 77–86 (2006)
7. Luyckx, K., Daelemans, W.: Shallow text analysis and machine learning for authorship attribution
8. Pavelec, D., Justino, E., Batista, L.V., Oliveira, L.S.: Author identification using writer-dependent and writer-independent strategies. In: Proceedings of the 2008 ACM symposium on Applied computing. pp. 414–418. SAC '08, ACM, New York, NY, USA (2008), http://doi.acm.org/10.1145/1363686.1363788
9. Peng, F., Schuurmans, D., Wang, S.: Augmenting naive bayes classifiers with statistical language models (2003)
10. Potthast, M., Eiselt, A., Barrón-Cedeño, A., Stein, B., Rosso, P.: Overview of the 3rd international competition on plagiarism detection. In: CLEF (Notebook Papers/Labs/Workshop) (2011)
11. Sidorov, G., Velasquez, F., Stamatatos, E., Gelbukh, A., Chanona-Hernández, L.: Syntactic dependency-based n-grams as classification features. In: Batyrshin, I., Mendoza-González, M. (eds.) Advances in Computational Intelligence, Lecture Notes in Computer Science, vol. 7630, pp. 1–11. Springer Berlin Heidelberg (2012), http://dx.doi.org/10.1007/978-3-642-37798-3_1
12. Sowa, J.F.: Conceptual structures: information processing in mind and machine. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA (1984)
13. de Vel, O., Anderson, A., Corney, M., Mohay, G.: Mining e-mail content for author identification forensics. SIGMOD RECORD 30, 55–64 (2001)
14. Yan, X., Han, J.: gspan: Graph-based substructure pattern mining (2002)