Evaluación de características para el proceso de atribución de autoría

Esteban Castillo, Darnes Vilariño, David Pinto, Maya Carrillo e Iván Olmos

Benemérita Universidad Autonóma de Puebla,
Facultad de Ciencias de la Computación,
Av. San Claudio y 14 Sur, Puebla, México
ecjbuap@gmail.com,{darnes,dpinto,cmaya, iolmos}@cs.buap.mx

Resumen La detección de autoría es una tarea de categorización que intenta descubrir el nombre del autor para un cierto documento anónimo dado como entrada. Para resolver esto, se necesita determinar primordialmente el conjunto de características que distinguen el modo de escritura de cada autor. En el presente trabajo se discuten los resultados obtenidos al evaluar diversas características para el proceso de atribución de autoría. El estudio se ha realizado utilizando tres corpora distintos: los dos primeros fueron proporcionados en el foro denominado "Uncovering Plagiarism, Authorship, and Social Software Misuse Lab (PAN 2011)", mientras que el tercero es un corpus balanceado que proviene de un trabajo altamente citado en el estado del arte, conteniendo documentos de noticias corporativas industriales. De acuerdo a los experimentos realizados, las características elegidas permiten representar documentos escritos por diferentes autores, de tal manera que un clasificador supervisado basado en máquinas de soporte vectorial puede asignar, en promedio, la autoría de un cierto documento anónimo con un 60 % de exactitud.

Keywords: Atribución de autoría, extracción de características, clasificación, máquinas de soporte vectorial

1. Introducción

En la forma más básica, el problema de atribución de autoría consiste en determinar de manera unívoca el autor correcto de un documento anónimo. En años recientes se ha incrementado el número de investigaciones en este sentido, y el problema ha sido enmarcado dentro del modelo o paradigma de categorización de textos, es decir, dicha problemática ha sido llamada de distintas maneras tales como adjudicación de textos, determinación de autoría, o como actualmente se le conoce atribución de autoría.

La búsqueda de características para representar a un autor específico supone una barrera que depende del tipo de lenguaje utilizado en cada texto. Si bien el lenguaje está limitado por el tema tratado o por el público a quién va dirigido (entre otros factores), el autor aun es libre de escribir como desea hacerlo, (lo cual

supone un obstáculo más en la detección de dicho autor), es decir, un autor es libre de hacer oraciones o párrafos de la longitud que desee, de usar el vocabulario y los signos de puntuación como lo considere conveniente, o de usar el tiempo verbal que más le agrade. Lo anterior nos lleva al hecho de que a pesar de haber muchos elementos impredecibles en la detección de autoría (como un conjunto de autores abierto), la búsqueda de características que cree o plasmen fielmente los principales rasgos de un autor, es una cuestión primordial en el proceso de entrenamiento sobre el mismo.

La presente investigación tiene como objetivo mostrar los resultados obtenidos al detectar los autores reales de un conjunto de documentos, considerando la representación tanto de los documentos de entrenamiento, como los de prueba utilizando 18 características seleccionadas (léxicas y sintácticas). El trabajo de investigación se desarolló usando tres corpora conformados de diferente manera. El objetivo fue observar el comportamiento de las características seleccionadas sobre colecciones de textos que no comparten el dominio ni el tipo de escritura.

Es por ello que nos hemos planteado la siguiente pregunta de investigación: En que medida es posible detectar los rasgos distintivos de cada autor, independientemente del corpus con que se trabaje?

El resto de este artículo está estructurado de la siguiente manera: En la sección 2 se discuten los resultados reportados hasta el momento en la literatura para darle solución a este tipo de problema. En la sección 3 se presentan las características a evaluar y el proceso de clasificación. Por último, se comenta la evaluación realizada y se dan las conclusiones obtenidas en esta investigación.

2. Trabajos previos

Las investigaciones mas relevantes realizadas hasta el momento se encaminan fundamentalmente hacia el tipo de características que realmente se deben utilizar, es decir, si deberían ser características léxicas, sintácticas o semánticas. Adicionalmente, ciertos resultados importantes se han desarrollado buscando la mejor forma de representar los documentos, considerando algunas de las características mencionadas con anterioridad. En este sentido, a continuación se presentan los trabajos más recientes con respecto a la extracción y uso de características en la representación de los elementos de escritura distintivos de un cierto autor.

Stamatatos [6], por ejemplo, presenta un estudio acerca de los elementos más importantes a tomar en cuenta en la tarea de atribución de autoría. El estudio es bastante completo y considera una revisión del estado del arte de este problema desde sus inicios hasta la actualidad, haciendo énfasis en la extracción de características y de cómo eśtas ayudan en el proceso de detección de un autor. En este trabajo, Stamatatos presenta los siguientes tipos de características como la base de los elementos más usados para representar el estilo: características estilométricas, léxicas, semánticas, de caracter y de estructura. El estudio muestra las dificultades y los retos aún sin resolver en el área de atribución de autoría.

En [2], se hace una revisión de las principales diferencias en los distintos tipos de características presentes en los textos. Se proponen también medidas para obtener el grado de acierto con respecto a las técnicas usadas para atribuir el autor de un documento anónimo dado.

Shlomo [1], por otro lado, presenta un resumen de la evaluación de la metodología aplicada para la atribución de autoría en el congreso PAN 2011 mostrando las distintas soluciones propuestas por los participantes del foro para resolver este problema. Dentro del análisis hecho en ese trabajo se pueden identificar cuatro clases de elementos fundamentales con respecto a las características usadas: Características derivadas del uso de las palabras en el texto (Palabras, N-gramas de palabras, Pronombres, Palabras del discurso, Contracciones y abreviaciones), Características a nivel de caracter (N-gramas de caracteres, Sufijos, Prefijos, Puntuación, Longitud de sentencia y palabras), Características basadas en el formato del texto (Ortografía, Orden de las estructuras gramaticales) y Características basadas en la sintaxis del texto (Partes del discurso, Sintaxis).

En [3] se plantea la idea de usar diferentes clasificadores para atribuir la autoría de un documento anónimo. Cada clasificador se entrena con una sola característica y posteriormente se establece un sistema de "votación", en donde un conjunto de características final es determinado como representativo de un autor si y solo si dicho conjunto presenta un comportamiento óptimo en los distintos clasificadores. Entre las características más importantes se usan las siguientes: Número de palabras, Número de líneas de texto, Número de caracteres en un documento, Número de sentencias, Número de bloques de texto, Número de palabras cerradas y Uso de las letras capitales. La relevancia de ese artículo radica en la combinación de distintos clasificadores para probar el peso y el grado de utilidad de un conjunto de características con respecto a un autor dado.

Si bien, existen otros trabajos en la literatura relacionados con la atribución de autoría, hasta donde sabemos, los anteriores resultan ser los más representativos.

En general, los resultados que se han obtenido hasta el momento muestran que las características seleccionadas por los autores depende mucho del corpus con el que se está trabajando. Es por ello que en la presente investigación se propone trabajar con 17 características sobre tres corpora que poseen estructuras estilométricas diferentes. El objetivo principal es analizar si las características propuestas logran detectar los rasgos distintivos de cada autor, independientemente de las estructuras textuales existentes en los diferentes corpora.

3. Extracción de características y proceso de clasificación

Una de las principales preocupaciones en la atribución de autoría es la búsqueda de todas aquellas propiedades cuantificables de un autor, que sean capaces de diferenciarlo de otros. A este tipo de elementos presentes en la mayoría de los textos se les llama características. Es de suma importancia el determinar las características adecuadas, y de preferencia óptimas que califiquen el estilo de escritura de cada autor.

3.1. Características escogidas

Para la extracción de las características representativas se consideró conservar intacto el texto de cada autor, en este caso, únicamente se eliminaron las etiquetas XML presentes en los textos. Se usa un corpus de entrenamiento que asocia a cada documento su correspondiente autor, con la finalidad de entrenar un modelo supervisado. La descripción del corpus de entrenamiento y del modelo de clasificación se explica más adelante. A continuación se presentan y describen las características a evaluar en este trabajo de investigación.

- Características a nivel de oración:
  - Palabras con frecuencia igual a 1. Son aquellas palabras que aparecen una sola vez en cada documento que pertenece al mismo autor.
  - Palabras con frecuencia igual a 2. Son aquellas palabras que aparecen dos veces en cada documento que pertenece al mismo autor.
  - Palabras más frecuentes. Las 10 palabras que aparecen con mayor frecuencia en los documentos de un autor dado.
  - Colocaciones. Los pares de palabras que siempre aparecen juntas en un documento dado (bigramas)[1].
  - Prefijos de palabras. La subcadena que antecede a la base léxica.
  - Sufijos de palabras. La subcadena que precede a la base léxica.
  - Trigramas de palabras. Todas las subsecuencias de tres palabras de cada documento.
  - Palabras cerradas. Las palabras con poco contenido semántico, como artículos, pronombres, preposiciones, etc.
  - Número de palabras por oración. Se cuantifica para cada documento el número de palabras que posee cada oración.
  - Número de oraciones. Se cuantifica cuantas oraciones posee cada documento.
  - Trigramas de las categorías gramaticales del documento. Todas las secuencias de 3 categorías que aparecen a lo largo de todo el documento.
- Características a nivel de caracter:
  - Combinación de vocales. Se elimina de cada palabra las consonantes y se considera la combinación de vocales restantes como una característica (se unen todas las vocales repetidas en cada combinación).
  - Permutación de vocales. Se elimina de cada palabra las consonantes y se consideran todas las combinaciones de vocales, cada combinación como una característica.
  - Letras del abecedario. Se considera el número de veces que aparece cada letra del abecedario en un documento dado.
  - Trigrama de caracter. Se cuantifican todas las subsecuencias de tres caracteres en un documento dado.
  - Número de caracteres por oración. Se cuantifica para cada documento el número de caracteres que posee cada oración.
  - Signos de puntuación. Se cuantifican todos los delimitadores de frases y párrafos que establecen una jerarquía sintáctica.

[1] Se utilizó el paquete NLTK para obtener estas colocaciones.

3.2. Método de clasificación

Sea $\langle x_1, x_2, \ldots, x_n \rangle$ el conjunto de características seleccionadas para representar los documentos. Cada documento es representado considerando la frecuencia de aparición de cada una de las características presentadas anteriormente, es decir, se usa el modelo de bolsa de palabras ponderadas para representar a cada documento[5].

Como modelo clasificador se utiliza una máquina de soporte vectorial (Support Vector Machine, por sus siglas en inglés SVM). SVM es un sistema de aprendizaje basado en el uso de un espacio de hipótesis de funciones lineales en un espacio de mayor dimensión inducido por un Kernel, en el cual las hipótesis son entrenadas por un algoritmo tomado de la teoría de optimización, el cual utiliza elementos de la teoria generalización. Debido a las limitaciones computacionales de las máquinas de aprendizaje lineal, éstas no pueden ser utilizadas en la mayoría de las aplicaciones del mundo real. La representación por medio del Kernel ofrece una solución alternativa a este problema, proyectando la información a un espacio de características de mayor dimensión, lo cual aumenta la capacidad computacional de las máquinas de aprendizaje lineal. Se mapea el espacio de entrada X a un nuevo espacio de características:

$$x = \{x_1, x_2, \ldots, x_n\} \rightarrow \phi(x) = \{\phi(x)_1, \phi(x)_2, \ldots, \phi(x)_n\} \tag{1}$$

Utilizando la función Kernel no es necesario calcular explícitamente el mapeo $\phi : X \rightarrow F$ para aprender en el espacio de características.

En el caso de esta investigación se utilizó como Kernel el mapeo polinomial que es un método muy popular para modelar funciones no lineales:

$$K(x, x) = (\langle x, x \rangle + C)^d \tag{2}$$

donde $C \in R$.

La SVM utilizada es la implementada en la plataforma de software para aprendizaje automático y minería de datos WEKA[7], ya que está disponible como software libre, es portable y contiene una extensa colección de técnicas para el preprocesamiento de datos y modelado.

4. Evaluación

Se desarrollaron 18 experimentos para evaluar las características de autoría planteadas anteriormente. En los 17 primeros experimentos se usa siempre una sola característica, y se representa a cada autor considerando la frecuencia ésta en cada uno de sus documentos. El último experimento considera entrenar el modelo de clasificación usando las seis mejores características de acuerdo a las evaluaciones obtenidas en los primeros 17 experimentos. El proceso de entrenamiento se llevó a cabo en los corpora que a continuación se describen:

4.1. Conjunto de datos

Se trabajó con tres conjuntos de datos en inglés, el primero y segundo fueron propuestos en el marco del congreso PAN 2011. Se trata de dos colecciones de textos desbalanceadas conteniendo chats de la Empresa Enron; el primero (denominado *Small*) tiene 26 autores diferentes, y el segundo (llamado *Large*) tiene 72 autores diferentes. El tercer conjunto de datos (llamado *C50*) fue creado por Stamatatos a partir de la clase CCAT perteneciente a la colección RCV1 [4]. C50 consiste de noticias corporativas industriales y es básicamente una colección de documentos balanceados y disjuntos con 50 autores diferentes. En la Tabla 1 se muestra el número de documentos usados tanto para la fase de entrenamiento como para la fase prueba.

Tabla 1. Número de documentos, para entrenamiento y prueba, asociado a cada conjunto de datos

Conjunto de datos	Fase de entrenamiento	Fase de Prueba
C50	2,500	2,500
PAN Small	3,001	518
PAN Large	9,337	1,298

4.2. Resultados obtenidos

Los resultados para cada uno de los primeros 17 experimentos se pueden observar en la Tabla 3. Entre las características que obtienen un mejor rendimiento, es decir, que lograron recuperar mejor los rasgos distintivos de cada autor se encuentran las siguientes (se muestran los resultados para cada corpus):

- Trigrama del texto categorizado (44.92 %, 30.97 % y 58.30 %)
- Palabras cerradas (29.96 %, 18.48 % y 64.47 %)
- Permutación de las vocales (37.96 %, 7.70 % y 83.59 %)
- Sufijos de palabras (33.80 %, 7.30 % y 42.08 %)
- Combinación de las vocales (24.64 %, 6.50 % y 77.41 %)
- Letras del abecedario (28.88 %, 15.17 % y 71.62 %)

Sin embargo, la selección de las mismas contempló no solamente el porcentaje de valores correctos, sino tambiíen el cubrimiento de los documentos que corresponden a los autores. El proceso de selección se llevó a cabo de la siguiente manera. Primeramente se eligió la característica que obtuvo el mejor rendimiento global en las tres colecciones (en este caso, el "trigrama de texto categorizado"). Posteriormente, se fueron agregando aquellas características con buen porcentaje de documentos clasificados correctamente, pero que además recuperan documentos que las características anteriores no habían recuperado.

Con estos resultados en mente, se decide entonces realizar un nuevo experimento considerando ahora la frecuencia de las seis características anteriormente comentadas. Los resultados obtenidos pueden verse en la Tabla 3.

Tabla 2. Resultados de la representación de las características individuales

Característica	**C50** (2500 docs) % (correctos)	**PAN Large** (1298 docs) % (correctos)	**PAN Small** (518 docs) % (correctos)
Palabras de frecuencia 1	11.16 % (279)	13.32 % (173)	24.90 % (129)
Palabras de frecuencia 2	11.80 % (295)	10.86 % (141)	20.07 % (104)
10 palabras más frecuentes	16.48 % (412)	10.86 % (141)	20.07 % (104)
Colocaciones	12.92 % (323)	17.48 % (227)	51.93 % (269)
Signos de puntuación	16.32 % (408)	16.79 % (218)	58.88 % (305)
Combinación de las vocales	24.64 % (616)	6.50 % (85)	77.41 % (401)
Permutación de las vocales	37.96 % (949)	7.70 % (100)	83.59 % (433)
Sufijos de las palabras	33.80 % (845)	7.30 % (96)	42.08 % (218)
Prefijos de las palabras	15.80 % (395)	15.79 % (205)	59.65 % (309)
Trigrama de palabras	17.00 % (425)	23.11 % (300)	40.34 % (209)
Trigrama de caracteres	12.04 % (301)	24.73 % (321)	32.81 % (170)
Trigrama de texto categorizado	44.92 % (1123)	30.97 % (402)	58.30 % (302)
Núm. de oraciones	13.80 %(345)	14.71 % (191)	30.69 % (159)
Núm. palabras por oración	15.16 % (379)	15.94 % (207)	15.05 % (78)
Núm. caracteres por oración	16.08 % (402)	14.40 % (187)	41.11 % (213)
Palabras cerradas	29.96 % (749)	18.48 % (240)	64.47 % (334)
Letras del abecedario	28.88 % (722)	15.17 % (197)	71.62 % (371)

El número de documentos recuperados correctamente por el modelo de clasificación que conjuga las seis características es significativamente mejor que cualquiera de los resultados obtenidos de manera individual. Es interesante observar que en el caso del corpus *Small* del PAN se obtiene un porcentaje muy alto, lo cual podría ser derivado del hecho de que este corpus contiene documentos con una longitud mucho mayor que los otros dos corpora. Esto llevaría a poder identificar más facilmente ciertos patrones de escritura del autor.

Si bien, los resultados para los corpora *C50* y *PAN Large* aun se encuentran alrededor del 50 %, consideramos que es posible mejorarlos. Una línea de investigación que estamos trabajando es usando representaciones alternativas, como por ejemplos los grafos, y usando algoritmos que reconocen automáticamente, ciertos patrones, en este caso, de escritura.

Tabla 3. Resultados de la combinación de características

Característica	**C50** (2500 docs) % (correctos)	**PAN Large** (1298 docs) % (correctos)	**PAN Small** (518 docs) % (correctos)
6 Mejores características	53.20 % (1330)	46.53 % (604)	91.31 % (473)

5. Conclusión

En este trabajo de investigación se reporta una evaluación de diferentes características usadas para el proceso de atribución de autoría. Se puede observar, que para uno de los corpora es posible obtener un porcentaje alto de documentos clasificados correctamente, lo cual apoya la teoría de la posibilidad de detectar los rasgos distintivos de un determinado estilo de escritura. Para el caso de los otros dos corpora, con las características estudiadas se ha logrado clasificar correctamente más del 53 % de los documentos en el conjunto de datos *C50*, y

el 46 % de los documentos para el conjunto de datos *PAN Large*. Esto indica que se debe hacer un estudio más detallado para capturar patrones de escritura distintivos en el caso de textos cortos.

Como trabajo a futuro, nos planteamos analizar si modificando la representación textual mediante grafos, es posible mejorar los resultados de clasificación.

Referencias

1. Argamon, S., Juola, P.: Overview of the international authorship identification competition at PAN-2011. In: Petras, V., Forner, P., Clough, P.D. (eds.) CLEF (Notebook Papers/Labs/Workshop) (2011)
2. Juola, P.: Authorship attribution. Found. Trends Inf. Retr. 1(3), 233–334 (Dec 2006), http://dx.doi.org/10.1561/1500000005
3. Kern, R., Seifert, C., Zechner, M., Granitzer, M.: Vote/Veto Meta-Classifier for Authorship Identification - Notebook for PAN at CLEF 2011. In: Petras, V., Forner, P., Clough, P.D. (eds.) CLEF (Notebook Papers/Labs/Workshop) (2011)
4. Lewis, D.D., Yang, Y., Rose, T.G., Li, F.: Rcv1: A new benchmark collection for text categorization research. J. Mach. Learn. Res. 5, 361–397 (Dec 2004), http://dl.acm.org/citation.cfm?id=1005332.1005345
5. Manning, C., Raghavan, P., Schtze, H.: Introduction to information retrieval. In: Same working notes. pp. 117–120 (2008)
6. Stamatatos, E.: A survey of modern authorship attribution methods. J. Am. Soc. Inf. Sci. Technol. 60(3), 538–556 (Mar 2009), http://dx.doi.org/10.1002/asi.v60:3
7. Witten, I.H., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques. The Morgan Kaufmann Series in Data Management Systems, Morgan Kaufmann Publishers, San Francisco, CA, 2nd edn. (2005)