Determinación de género y edad en blogs en español mediante enfoques basados en perfil

Dario G. Funez, Leticia C. Cagnina y Marcelo L. Errecalde

Laboratorio de Investigación y Desarrollo en Inteligencia Computacional
Facultad de Ciencias Físico, Matemáticas y Naturales,
Universidad Nacional de San Luis - Ejército de los Andes 950
(D5700HHW) - San Luis - Argentina Tel: (0266) 4420823 / Fax: (0266) 4430224
e-mail: {dgfunez, lcagnina, merreca}@unsl.edu.ar

Resumen La determinación del perfil del autor (desconocido) de un documento permite identificar características como el género (sexo) y edad de dicho autor, en base al estilo de escritura y las palabras presentes en el documento. Esta tarea está creciendo en importancia en diferentes áreas de investigación, especialmente en idiomas como el español donde existen pocos trabajos realizados hasta el presente. En este trabajo se presentan los resultados obtenidos en la determinación de género y edad en blogs en español mediante *enfoques basados en perfil*, un tipo de técnica que ha sido aplicado exitosamente en tareas de atribución de autoría. Los resultados obtenidos muestran la viabilidad de la aplicación de este enfoque en la determinación del perfil del autor de un documento y permiten identificar aspectos que necesitan ser mejorados en el futuro.

Palabras Claves: categorización de documentos, minería de textos, perfiles de autores, enfoques basados en perfiles

1. Introducción

El uso creciente de redes sociales como `Facebook` y `MySpace`, sitios de microblogging como `Twitter` y las innumerables facilidades de chats disponibles hoy en día han hecho accesible mucha información provista por personas de diferentes edades, género, condición social, etc. Dicha información puede ser utilizada para inferir datos importantes del perfil del autor de un texto como su personalidad, demografía y antecedentes culturales [2]. La determinación del perfil del autor de un documento (en inglés *author profiling*) es la tarea de distinguir entre clases de autores en base a la forma del lenguaje compartido por un grupo social particular. Esto puede involucrar la identificación de diversos aspectos del perfil de una persona tales como el *género* (femenino vs masculino), *edad* (de acuerdo a distintos grupos etarios), *lenguaje nativo* y *tipo de personalidad*.

La actividad de recabar información del perfil del autor de un documento es un problema de interés creciente en áreas como seguridad y anti-terrorismo, marketing y diversas disciplinas forenses. En el caso de marketing, es evidente

el beneficio que puede obtenerse del empleo de los comentarios de clientes en blogs para determinar la demografía de la gente que gusta o no de determinados productos [1]. Asimismo, este tipo de técnicas también pueden tener un impacto importante en problemas forenses de abordaje más reciente como la detección automática de depredadores sexuales en la Web [7].

Por otra parte, en problemas de *atribución de autoría* [10] han ganado cada vez más relevancia los enfoques *basados en perfiles* [3,4,5,6] planteándose como alternativas interesantes a los enfoques clásicos de categorización de textos *basados en instancias* [10] debido a diversas ventajas como su facilidad de implementación, eficiencia de aplicación, escalabilidad y la representación explícita de información relevante sobre un autor.

En este trabajo analizamos si los enfoques basados en perfiles son adecuados para la determinación de la edad y el género de documentos de blogs en idioma español. Con respecto a la edad, se consideran posts de 3 grupos etarios distintos formados por adolescentes entre 13 y 17 años, jóvenes entre 23 y 27 años y adultos entre 33 y 47 años. El género por su parte puede ser femenino o masculino.

Estos enfoques fueron probados con la versión en español del corpus de entrenamiento del *PAN-PC-2013* [1] por ser el único disponible para experimentación a la fecha. Los resultados obtenidos sustentan la viabilidad de este tipo de técnicas y son un punto de comienzo para mejorar su desempeño en trabajos futuros.

El resto de este trabajo se organiza de la siguiente manera. En la Sección 2 se explican los principales conceptos vinculados a los enfoques basados en perfiles. La Sección 3 detalla los principales aspectos de la tarea abordada y cómo los enfoques basados en perfiles se ajustan a su resolución. La Sección 4 describe el trabajo experimental y el análisis de los resultados obtenidos. Finalmente, en la Sección 5 se exponen las conclusiones de nuestro estudio y se proponen trabajos futuros para mejorar esta propuesta.

2. Enfoques basados en perfiles

Los *enfoques basados en perfiles* constituyen uno de los enfoques principales hoy en día para abordar problemas de *atribución de autoría* (AA) [10]. En un problema de AA típico, un texto de autoría desconocida es asignado a un autor candidato, dado un conjunto de autores candidatos para los cuales se tienen disponibles textos de muestra de autoría indiscutida. Desde un punto de vista del aprendizaje automático (*machine learning*), esto puede ser visualizado como una tarea de categorización de texto de múltiples clases y único rótulo (*multi-class single-label*). En este contexto, para cada clase (autor) se construye un perfil de autor que contiene información recuperada de un conjunto de documentos escritos por el mismo [4]. En la la parte izquierda de la Figura 1 se muestra gráficamente el proceso de generación de los perfiles de cada autor.

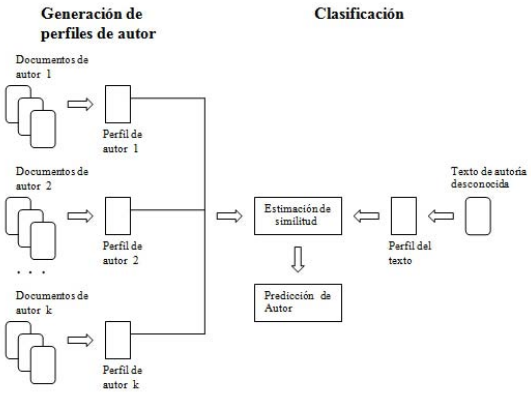

Figura 1: Atribución de autoría basada en perfil: generación de perfiles de autor (izquierda) y clasificación de un documento de prueba (derecha).

Las características a recuperar de un texto pueden estar basadas en el *estilo* de escritura o en el *contenido* del texto.

- *Características basadas en estilo*: extraen de los documentos medidas estilográficas como la frecuencia de determinadas clases de palabras como pronombres, artículos, preposiciones, cantidad de hipervínculos, promedio de palabras en un post etc. [9]. Estas características usualmente varían dependiendo del género y la edad. Por ejemplo, las mujeres en los blogs suelen utilizar más pronombres y palabras de aprobación y negación, reduciéndose esta tendencia en edades más avanzadas.
  Una de las características más utilizadas para capturar aspectos de estilo son las frecuencias de n-gramas de caracteres. Los n-gramas son subcadenas de n caracteres consecutivos [3] siendo común el uso de $n = 3$, 4 y 5. En el inglés por ejemplo, el uso de tri-gramas de caracteres demostró ser efectivo para capturar la frecuencia de adverbios, información contextual, etc.
- *Características basadas en contenido*: consideran las palabras que pertenecen a temáticas particulares [9]. En esta categoría también existen diferencias en su uso dependiendo del género del autor. Por ejemplo, las escritoras tienden a utilizar con más frecuencia palabras relacionadas con lo personal como por ejemplo “shopping”, “madre”, etc. En cambio, los escritores se interesan más en temas relacionados a política y tecnología. Los adolescentes escriben sobre sus amigos, estados de humor y temas relacionados al colegio. A edades mayores crece su interés por el casamiento, la política y temas financieros.

Para obtener el perfil de un autor se considera un conjunto de documentos de su autoría y se extraen de él un conjunto de L características. Por ejemplo, para el caso en el que se seleccione como característica los tri-gramas (subcadenas de 3 caracteres), se deben obtener todos los tri-gramas de cada documento y se los ordena por su frecuencia. Luego, los L tri-gramas más frecuentes constituirán el perfil. Para poder clasificar un documento a un autor, se necesita generar el

perfil del documento y luego, utilizando una medida de distancia (o de similitud), se determina si existen similitudes entre el perfil del documento y el perfil de cada autor [6]. El autor cuyo perfil sea el más "cercano" al perfil del documento, será el que se retorne en el proceso de clasificación, como se muestra en la parte derecha de la Figura 1. Algunas de las medidas de distancia/similitud utilizadas en los enfoques basados en perfil son las siguientes:

- *Keselj's Relative Distance (KRD)*: esta medida, referenciada en algunos trabajos como *CNG*, mide la distancia K entre dos perfiles P_1 y P_2 como

$$K = \sum_{x \in X_{P_1} \cup X_{P_2}} \left(\frac{2 \times (P_1(x) - P_2(x))}{P_1(x) + P_2(x)} \right)^2$$

  donde $P_i(x)$ es la frecuencia del término x en el perfil P_i, y X_{P_i} es el conjunto de todos los términos que ocurren en el perfil P_i.
- *Simplified Profile Intersection (SPI)*: Esta medida de similitud es una versión simplificada de la anterior, que sólo toma en cuenta la cantidad de características que pertenecen a ambos perfiles [5].
- *Out of Place (OOP)*: Estima la diferencia posicional de cada característica en los perfiles a comparar y la medida es la suma de todas las diferencias posicionales para todas las características de los perfiles [3].

Un inconveniente con las medidas enunciadas previamente es que muchos de los términos que ocurren en los perfiles de los autores son términos utilizados muy frecuentemente en el lenguaje apareciendo en consecuencia en todos los perfiles y por lo tanto aportando poca información discriminativa entre un perfil y otro. Esta observación, ha conducido a nuevas propuestas como el *recentrado de perfiles locales* que se han aplicado recientemente con éxito en tareas de AA [6] y consideramos que pueden ser relevantes en el problema abordado en el presente trabajo, por lo que será descripto en forma más detallada a continuación.

El *recentrado de perfiles locales* (RPL) crea perfiles priorizando aquellas características que son usadas en forma diferencial respecto al uso del lenguaje habitual [6]. Para medir el uso del lenguaje habitual, el RPL utiliza un *perfil del lenguaje* que aproxima el uso real de las características en el lenguaje midiendo las frecuencias de ocurrencia en todos los documentos del conjunto de entrenamiento. Si bien ésta es sólo una de las posibles implementaciones alternativas para usar como perfil del lenguaje, en [6] ha demostrado ser bastante efectiva.

De esta forma, para construir el perfil de un autor en RPL se utilizan las L primeras características, ordenadas (en forma decreciente) por el valor absoluto de la diferencia entre su valor de frecuencia de uso en el perfil del autor y su frecuencia de uso en el perfil del lenguaje.

El algoritmo RPL no sólo realiza un "recentrado" de los valores de los perfiles de cada autor respecto al perfil del lenguaje; también utiliza una función de distancia específica que se define como:

$$d(f_1, f_2) = \sum_{x \in profiles} \frac{(f_1(x) - E(x)) \times (f_2(x) - E(x))}{\|f_1(x) - E(x)\| \times \|f_2(x) - E(x)\|}$$

donde f_1 y f_2 son los perfiles a ser comparados, $f_i(x)$ es la frecuencia normalizada de la característica x en el perfil f_i, E es el modelo del lenguaje y el término $profiles$ denota el conjunto de las características que se encuentran en las primeras L posiciones de f_1 o f_2. Como se puede observar, esta medida no es más que la distancia coseno de las cuentas recentradas de ambos perfiles. Esto implica que el perfil del documento a ser clasificado también deba ser recentrado bajo este esquema.

3. Descripción del sistema clasificador

Nuestro estudio en este trabajo se enfoca en analizar la viabilidad del uso de enfoques basados en perfil, los cuales se han desempeñado con gran éxito en tareas de AA, en tareas de determinación del perfil de un autor de documentos en español. Específicamente, esta tarea consistirá en predecir el género del autor (femenino o masculino) y la edad del mismo (grupo de los 10's, 20's o 30's). El grupo de las edades de "10" comprende las edades entre 13-17 años, el de 20 entre 23-27 y el de 30 entre 33-47 años [1].

Como se puede observar, un perfil para cada una de estas clases no representará un autor particular sino una *clase de autores* de acuerdo a su grupo etario o el sexo de la persona. Así, por ejemplo, un perfil que se construya con documentos rotulados como "20" representará características de autores cuya edad oscila entre los 23 y 27 años en lugar de características de un autor particular.

Para usar un enfoque basado en perfil para determinar el género y la edad de un autor, se deben definir dos aspectos principales. En primer lugar, si los perfiles que se utilizarán consideran a cada una de estas subtareas (determinar el sexo y la edad) como dos tareas separadas o no. En el primer caso, se definirán perfiles separados para la determinación del género (uno para femenino y otro para masculino) y para la determinación de la edad (un perfil por cada grupo etario, tres perfiles en total). En el segundo enfoque, se considera que el hecho de tratar la edad y el género en forma conjunta y simultánea puede ser beneficioso y se definirá un perfil para cada una de las 6 posibles combinaciones de las categorías consideradas: "femenino-10", "masculino-10", "femenino-20", etc.

El otro aspecto a definir es el tipo de característica a utilizarse en los perfiles: palabras completas, n-gramas de caracteres, características estilográficas, etc. Los experimentos preliminares mostraron que el uso de palabras completas en lugar de n-gramas de caracteres y la utilización de perfiles de categorías combinadas ("género-edad") producen mejores resultados por lo que se describirá en el resto de esta sección cómo se implementó este enfoque.

El sistema completo se implementó en dos etapas. En la primera se generaron los perfiles para cada una de las 6 categorías combinadas consideradas: *masc-10, fem-10, masc-20, fem-20, masc-30* y *fem-30*. Para implementar el enfoque recentrado se debió obtener además el perfil del lenguaje. La característica que se utilizó para generar los perfiles es la de palabras completas, la cual demostró ser superior a distintas variantes de n-gramas de caracteres como las de 3-gramas, 4-gramas, 5-gramas y combinaciones de ellas (3-gramas y 4-gramas, 4-gramas y

5-gramas). Los documentos empleados para construir los perfiles, forman parte del corpus español de entrenamiento de la competencia *PAN-PC-2013* [1].

El perfil del lenguaje se obtuvo realizando los siguientes pasos en secuencia considerando todo el conjunto de entrenamiento:

1. *Generación de perfil por cada documento*: Para cada documento del conjunto de entrenamiento se obtiene su perfil de palabras, donde cada palabra tiene como valor asociado su frecuencia de ocurrencia en el documento. Estos perfiles fueron generados con la librería Morphadorner[1].
2. *Unificación de los perfiles en un único perfil*: En esta tarea se concatenan todos los perfiles obtenidos en el paso anterior, obteniéndose un perfil que repite una característica si ésta ya aparece en otro documento.
3. *Eliminación de entradas repetidas*: El perfil del lenguaje se logra sustituyendo las entradas de las palabras repetidas en una única entrada, con un valor que es el total de sumar todos los valores de esas entradas, normalizados por el número de documentos. De esta manera se calcula la acumulación de todos los valores de una característica.

Para obtener el perfil de una categoría se aplican las mismas tareas que para obtener el perfil del lenguaje, pero restringiéndose a los documentos propios de esa categoría. Si se trabaja con perfiles recentrados, se debe recentrar el valor de cada característica con los valores del perfil del lenguaje, y posteriormente se ordena considerando el valor absoluto del valor recentrado.

La segunda etapa del sistema clasificador consiste básicamente en la implementación del proceso de clasificación de un documento de test arbitrario d_t utilizando los perfiles obtenidos en la primera etapa. El clasificador recibe como parámetro de entrada un archivo *xml* que tiene el formato de una conversación en un blog. Este archivo, en los enfoques basados en perfiles sin recentrado recibe un preprocesamiento básico y se genera su perfil de documento, clasificándose de acuerdo al esquema mostrado en la parte derecha de la Figura 1, usando el perfil de categoría más cercano de acuerdo a alguna de las funciones de similitud/distancia descriptas en la Sección 2 (KRD, SPI u OOP).

En un método como RPL que utiliza recentrado de perfiles, el procedimiento de clasificación es un poco más complejo como se muestra en la Figura 2 y se describe a continuación:

- *Preprocesamiento del archivo de entrada*: Se eliminan los tags de las conversaciones y se sustituyen las imágenes por tres caracteres (IMG) para no perder información sobre el contenido en el documento. También los dígitos fueron reemplazados por un caracter especial definiendo patrones a reemplazar en el documento.
- *Generación del perfil del documento ordenado por frecuencias*: En este módulo se consigue el perfil de palabras del documento a clasificar.
- *Recentrado respecto del perfil del lenguaje*: En esta tarea se recentra el valor de todas las palabras del perfil del documento.

[1] *Morphadorner* es una librería escrita en lenguaje Java, de acceso libre para PLN y suministrada por la Universidad de Northwestern.

- *Chequeo por similitud (o disimilitud) con los perfiles de cada categoría*: Se compara el perfil del documento con el de cada categoría, retornándose el rótulo de aquella que es más cercana (menos distante).

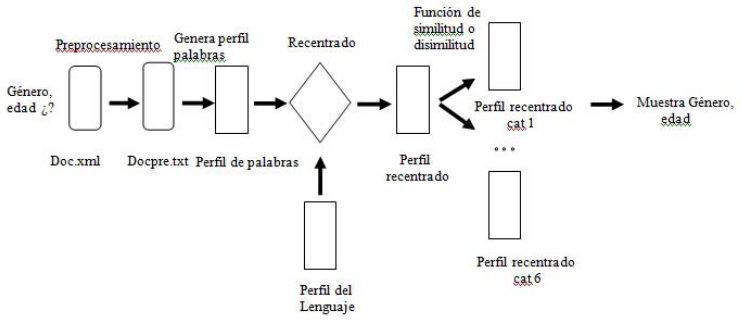

Figura 2: Diagrama del sistema clasificador utizando recentrado.

4. Experimentos

Para evaluar el desempeño de los distintos enfoques de clasificación basados en perfiles se obtuvieron dos conjuntos de prueba extraídos del corpus de entrenamiento de la competencia *PAN-PC-2013* [1]. El primer conjunto se utilizó para evaluar el género y está compuesto de 1000 documentos, con 500 archivos femeninos y 500 masculinos. En cada categoría, 166 archivos son de tamaño pequeño, 166 medianos y 166 grandes (considerando el tamaño en bytes del documento). El criterio utilizado para generar un conjunto de prueba con estas características, fue el de tener una colección con un número balanceado de documentos representativos de las distintas categorías, tamaños de documentos, etc. El otro conjunto de prueba se utilizó para evaluar la edad con un total de 1000 archivos divididos en tres categorías (10s, 20s, 30s) de 333 archivos aproximadamente cada uno, cada categoría también balanceada en el número de documentos de distintos tamaño.

Para que el clasificador sea aceptable, debe superar el "baseline" de 0.50 para el género y 0.33 para la edad, cuyos valores corresponden a una selección equiprobable entre las categorías disponibles y que también ha sido utilizado en la competencia. Los enfoques basados en perfiles que consideraremos en nuestro trabajo son los que construyen los perfiles en forma "clásica" tomando los L términos (palabras completas) más frecuentes. Utilizaremos en este caso, como medidas de similitud/distancia las siguientes 3 medidas que se describieron en la Sección 2: KRD, SPI y OOP. También analizaremos el desempeño de un enfoque que trabaja con perfiles recentrados como es el caso del algoritmo RPL. Respecto a este último punto, una pregunta que surgió al realizar este estudio, fue cual sería el impacto de trabajar con un perfil recentrado (como en RPL) pero en lugar de utilizar la función de distancia (coseno) propia de RPL medir la similitud/distancia entre los perfiles utilizando algunas de las funciones que utilizamos previamente en los enfoques no recentrados (SPI u OOP). Al uso de estas

funciones de distancia combinado con el recentrado del perfil los denotaremos SPI^{re} y OOP^{re} respectivamente.

En la Tabla 1 se muestran los resultados de la experimentación para el género en español, tomando como referencia el porcentaje de aciertos (clasificaciones correctas, en inglés *accuracy*) obtenido con los distintos enfoques basados en perfiles. Para cada uno de ellos, se especifica el resultado obtenido con distintos valores de L (tamaño del perfil) desde 200 hasta 8000. Como se puede observar, un enfoque basado en recentrado como RPL no obtiene resultados superiores al *baseline* (0.5) para ninguno de los valores de L considerados. Algo similar se observa cuando se usan perfiles recentrados con OOP como función de distancia (enfoque OOP^{re}). En este sentido, el único caso en que el uso de perfiles recentrados obtiene resultados por encima del baseline es el de SPI^{re}.

Respecto a los enfoques que utilizan los perfiles "clásicos" (KRD, SPI y OOP), estos obtienen mejores resultados superando al baseline con todos los valores de L considerados. En algunos casos incluso, se obtienen valores cercanos o levemente superiores a 0.6 que son comparables a valores preliminares reportados en tareas similares de categorización de género en blogs [1]. Es importante remarcar que con algunas medidas de distancia como SPI, el incremento progresivo del valor de L por encima del tamaño máximo reportado en este trabajo (8000), puede generar mejores porcentajes de aciertos con el conjunto de prueba considerado. Así por ejemplo, usando SPI con $L = 40000$ se logra una "accuracy" de 0.673. Se debe tener en cuenta sin embargo, que esto se logra a costa de un incremento significativo del tiempo de CPU requerido en tiempo de prueba, los cuales pueden ser inaceptables cuando se deben categorizar decenas de miles de documentos. Por otra parte, este incremento del L para mejorar la accuracy, también podría generar un efecto de "sobreajuste" (en inglés *overfitting*) a las características particulares del conjunto de prueba utilizado. Elegir un valor de L adecuado para lograr un balance correcto entre tiempos de clasificación aceptables, buen porcentaje de aciertos y evitar el efecto de sobreajuste es un factor importante que será abordado en trabajos futuros.

Respecto a la determinación de la edad del autor en blogs en español, los resultados de la Tabla 2 confirman lo observado previamente respecto a la baja efectividad del uso de perfiles recentrados en este tipo de tareas. En los 3 casos que utilizan perfiles recentrados (RPL, SPI^{re} y OOP^{re}) ninguno de los valores de L utilizados permitió superar el baseline de 0.33 para esta tarea.

También aquí, KRD, SPI y OOP superan ampliamente el baseline para esta tarea, lográndose en el caso de SPI los mejores valores de "accuracy" (0.565) con $L = 8000$. Nuevamente se debe remarcar en este caso, que un incremento de los valores de L puede generar mejores resultados (0.641 con SPI y $L = 40000$) siendo válidas las mismas consideraciones respecto al costo de clasificación y sobreajuste que se realizaron para el caso del género.

Tabla 1. *Accuracy* para determinación de género en español. *Baseline* = 0.5.

L	KRD	SPI	OOP	RPL	SPI^{re}	OOP^{re}
200	0.544	0.532	0.57	0.473	0.502	0.475
500	0.564	0.546	0.559	0.471	0.501	0.476
1000	0.58	0.553	0.58	0.471	0.51	0.472
2000	0.589	0.58	0.602	0.469	0.531	0.475
3000	0.562	0.571	0.56	0.472	0.526	0.483
4000	0.572	0.572	0.58	0.472	0.536	0.485
6000	0.599	0.581	0.593	0.472	0.538	0.488
8000	0.572	0.57	0.568	0.472	0.54	0.487

Tabla 2. *Accuracy* para determinación de edad en español. *Baseline* = 0.33.

L	KRD	SPI	OOP	RPL	SPI^{re}	OOP^{re}
200	0.435	0.428	0.432	0.313	0.267	0.33
500	0.436	0.445	0.44	0.313	0.264	0.37
1000	0.464	0.488	0.464	0.313	0,267	0.345
2000	0.493	0.497	0.496	0.313	0.267	0.318
3000	0.476	0.526	0.492	0.313	0.267	0.317
4000	0.483	0.531	0.483	0.313	0.268	0.314
6000	0.489	0.544	0.491	0.313	0.269	0.310
8000	0.504	0.565	0.493	0.313	0.27	0.311

5. Conclusiones y Futuras Extensiones

En este trabajo se analizó la viabilidad del uso de enfoques basados en perfiles para la determinación del género y la edad en blogs en español. De acuerdo a nuestro conocimiento, esta es la primera vez que se realiza un estudio de esta naturaleza. De los estudios preliminares, se pudo observar que para el español, el uso de palabras completas es más efectivo que los n-gramas de caracteres y que la determinación conjunta y simultánea de la edad y el género es más efectiva que considerarlas como tareas separadas. Observaciones similares se han realizado para el lenguaje holandés utilizando métodos de clasificación basados en instancias como SVM [8].

Otra observación interesante es que a pesar de su atractivo y efectividad en tareas de AA, el recentrado de perfiles no parece obtener resultados competitivos en esta tarea. Como trabajo a futuro, se planea un análisis más detallado para determinar las causas de este bajo desempeño. Sin embargo, métodos basados en perfiles clásicos como KRD, SPI y OOP obtienen resultados competitivos, comparables a los de otros enfoques más complejos y costosos. En este sentido, SPI a pesar de su simplicidad, ha mostrado resultados prometedores, en particular cuando se incrementa el valor de L para el perfil. Encontrar un valor de L que combine de manera adecuada la precisión en la predicción, bajo costo de clasificación y evite el sobreajuste es un tema de estudio futuro a desarrollarse

cuando se disponga de un conjunto más grande de datos de prueba, como los que serán liberados próximamente como parte del "test set" de la competencia *PAN-PC-2013*.

Es importante observar que más allá de haberse mostrado la viabilidad del uso de este tipo de enfoques en esta clase de problemas, existe un amplio campo para mejorar estos enfoques como pueden ser una selección más cuidadosa de los documentos usados para generar los perfiles de las categorías, el uso de características más informativas para representar los documentos y la combinación de este tipo de métodos con métodos basados en instancia.

Referencias

1. 9th evaluation lab on uncovering plagiarism, authorship, and social software misuse (PAN 2013). `http://pan.webis.de/`, 2013.
2. Shlomo Argamon, Moshe Koppel, James W. Pennebaker, and Jonathan Schler. Automatically profiling the author of an anonymous text. *Communications of the ACM*, 52:119–123, 2009.
3. William B. Cavnar and John M. Trenkle. N-gram-based text categorization. In *Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval*, pages 161–175, 1994.
4. Hugo Jair Escalante, Manuel Montes y Gómez, and Thamar Solorio. A weighted profile intersection measure for profile-based authorship attribution. In *Proceedings of MICAI 2011*, volume 7094, pages 232–243, 2011.
5. Georgia Frantzeskou, Efstathios Stamatatos, Stefanos Gritzalis, and Sokratis Katsikas. Source code author identification based on n-gram author profiles. In *Artificial Intelligence Applications and Innovations*, volume 204 of *IFIP*, pages 508–515. Springer US, 2006.
6. Robert Layton, Paul Watters, and Richard Dazeley. Recentred local profiles for authorship attribution. *Natural Language Engineering*, 18:293–312, 2012.
7. India Mcghee, Jennifer Bayzick, April Kontostathis, Lynne Edwards, Alexandra Mcbride, and Emma Jakubowski. Learning to identify internet sexual predation. *International Journal of Electronic Commerce*, 15(3):103–122, April 2011.
8. Claudia Peersman, Walter Daelemans, and Leona Van Vaerenbergh. Predicting age and gender in online social networks. In *Proceedings of the 3rd international workshop on Search and mining user-generated contents*, SMUC '11, pages 37–44, New York, NY, USA, 2011. ACM.
9. Jonathan Schler, Moshe Koppel, Shlomo Argamon, and James W. Pennebaker. Effects of age and gender on blogging. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, pages 199–205, 2006.
10. Efstathios Stamatatos. A survey of modern authorship attribution methods. *Journal of the American Society For Information Science and Technology*, 60(3):538–556, 2009.