Procesamiento del Lenguaje Natural, núm. 43 (2009), pp. 225-233 recibido 1-05-2009; aceptado 5-06-2009

Clasificación de Páginas Web con Anotaciones Sociales

Web Page Classification with Social Annotations

Arkaitz Zubiaga, Raquel Martínez, Víctor Fresno
Universidad Nacional de Educación a Distancia
C/Juan del Rosal, 16, 20840 Madrid
{azubiaga, raquel, vfresno}@lsi.uned.es

Resumen: Las anotaciones generadas por usuarios en sistemas de marcadores sociales pueden proveer metadatos interesantes y muy útiles para la clasificación de páginas web. Estas anotaciones incluyen diversos tipos de información, como etiquetas y comentarios. No obstante, cada tipo de anotación tiene una naturaleza y un nivel de popularidad diferente. En este trabajo, analizamos y evaluamos la utilidad de cada una de estas anotaciones sociales para clasificar páginas web sobre una taxonomía como la del Open Directory Project. Las comparamos por separado a la clasificación basada en contenido, y también las combinamos. Nuestros experimentos muestran resultados prometedores con la utilización de anotaciones sociales para este propósito. Y además indican que su combinación con el contenido textual mejora el rendimiento de la clasificación.
Palabras clave: clasificación de páginas web, anotaciones sociales, marcadores sociales

Abstract: User-generated annotations on social bookmarking sites can provide interesting and promising metadata for web page classification. These annotations include diverse types of information, such as tags and comments. Nonetheless, each kind of annotation has a different nature and popularity level. In this work, we analyze and evaluate the usefulness of each of these social annotations to classify web pages over a taxonomy like that by the Open Directory Project. We compare them separately to the content-based classification, and also combine the different types of data. Our experiments show encouraging results with the use of social annotations for this purpose, and we found that combining these metadata with web page content improves even more the classifier's performance.
Keywords: web page classification, social annotations, social bookmarking

1. Introducción

Los sistemas de marcadores sociales como Delicious[1], StumbleUpon[2] y Diigo[3] permiten a sus usuarios describir contenidos web, anotándolos de forma colaborativa y conjunta. Estos datos pueden generarse de diferentes maneras: añadiendo etiquetas como descripción temática, haciendo valoraciones subjetivas o personales, añadiendo textos descriptivos, etc. Como resultado, se genera un gran volumen de metadatos que facilita la posterior recuperación y navegación sobre los documentos web.

El impacto de las anotaciones sociales ha sido explorado en múltiples contextos de la recuperación de información y la web, encontrándose estudios que analizan sus aspectos semánticos y el porqué de su popularidad y éxito (Noll y Meinel, 2008b). No obstante, muchas de las características de las anotaciones sociales están aún por explorar. Bao et al. (2007) descubrieron que las etiquetas proveen un resumen multifacético de los documentos web, y Noll y Meinel (2007) sugieren que las etiquetas pueden resultar apropiadas para tareas de clasificación. Estos autores indican que las etiquetas son cualitativamente diferentes al contenido. Otros trabajos han utilizado etiquetas para tareas de clustering (Ramage et al., 2009). Todos estos artículos apuntan a una pregunta: ¿Cómo pueden uti-

[1] http://delicious.com
[2] http://www.stumbleupon.com
[3] http://www.diigo.com

ISSN: 1135-5948

lizarse las etiquetas para mejorar las tareas de clustering, clasificación y recuperación de información? En este trabajo, ofrecemos una aproximación en lo que a clasificación se refiere.

La clasificación de páginas web se puede definir como la tarea de organizar documentos web en una serie de categorías predefinidas, problema que generalmente resulta ser multiclase. Las aproximaciones para clasificación de páginas web se diferencian de las de clasificación de textos en que aquéllas pueden hacer uso de características propias de la web, como el marcado HTML, hiperenlaces, analisis visual, etc. Los sistemas de etiquetado social proveen un nuevo tipo de información que puede ayudar en esta tarea.

En este artículo, estudiamos cómo las anotaciones de los usuarios de los marcadores sociales pueden ser utilizadas como datos complementarios al contenido textual de las páginas web para mejorar su clasificación. Consideramos como *baseline* la clasificación basada en contenido textual, y la comparamos con la clasificación basada en diferentes tipos de anotaciones sociales, tanto por separado como de forma combinada. No tenemos en cuenta el marcado HTML, hiperenlaces ni otra información estructural; nos centramos en el uso de contenido textual y anotaciones sociales.

El resto del artículo está organizado de la siguiente manera. La sección 2 presenta una perspectiva general de las anotaciones sociales, exponiendo los diferentes tipos y sus características. El trabajo previo en el área, y su relación con el nuestro se explica en la sección 3. Después, en la sección 4, se describe el proceso de generación de la colección de datos utilizada para los experimentos. En la sección 5, se explica el funcionamento de las máquinas de vectores de soporte. Continuamos con los detalles de la experimentación, presentando su configuración y analizando las diferentes representaciones utilizadas en la sección 6, mientras que la sección 7 muestra cómo combinamos los clasificadores y los resultados que estas combinaciones ofrecen. Finalmente, se presentan las conclusiones y el trabajo futuro en la sección 8.

2. Anotaciones Sociales

Los marcadores sociales permiten a los usuarios guardar y anotar sus páginas web preferidas. Estas anotaciones se realizan de forma colaborativa, lo que posibilita la acumulación de un gran volumen de metadatos para cada página web. Entrando en detalle en análisis de estos metadatos, pueden definirse diferentes tipos de anotaciones sociales:

Etiquetas (Tags): las palabras clave que definen y caracterizan una página web. A diferencia de los sistemas de etiquetado simple, donde únicamente un usuario anota un recurso (generalmente, el propio autor), un sistema de etiquetado colaborativo acumula las *etiquetas* de diferentes usuarios para un mismo recurso. Formalmente, cada usuario u_i puede guardar un recurso i_j con un conjunto de *etiquetas* $T_{ij} = \{t_1, ..., t_p\}$, con una cantidad variable p de etiquetas. Después de que k usuarios guardan el recurso i_j, se define como un conjunto de *etiquetas* con pesos $T_j = \{w_1t_1, ..., w_nt_n\}$, donde $w_1, ..., w_n \leq k$. Las *etiquetas* son el tipo de anotación más común, y está disponible en casi todos los sistemas de marcadores sociales.

Notas (Notes): los textos que describen el contenido de una página web. Junto a las *etiquetas*, las *notas* son el tipo de anotación más extendido en los marcadores sociales. Delicious, por ejemplo, ofrece la posibilidad de añadir *notas* además de *etiquetas*.

Destacados (Highlights): los usuarios pueden seleccionar la parte que consideran más relevante en la página web que guardan. Un servicio que utiliza *destacados* es Diigo.

Críticas (Reviews): texto en el que se valora una página web. Aunque a priori este tipo de anotaciones es subjetivo, los usuarios tienden a mezclar textos descriptivos junto con opiniones. StumbleUpon es uno de los sitios que ofrece esta característica.

Valoraciones (Ratings): puntuaciones, generalmente de 1 a 5, indicado el grado en el que a los usuarios les agrada o desagrada una página web. Como resultado, se ofrece un valor medio. Está disponible también en StumbleUpon.

La mayoría de marcadores sociales no establece restricciones para la selección de términos como *etiquetas* y, por tanto, los usuarios tienden a asignar *etiquetas* con diversos grados de especificidad y naturaleza. Éste ha sido uno de los motivos del éxito de las *etiquetas*. Sin embargo, aunque el vocabulario sea abierto, las *etiquetas* más populares tienden a ser más utilizadas por los usuarios (Noll y Meinel, 2007).

Finalmente, aunque la mayoría de las

anotaciones descritas en esta sección parecen prometedoras para tareas de clasificación temática de páginas web, es obvio que las *valoraciones* no ofrecen información temática. Por este motivo, hemos basado este estudio en todas las anotaciones sociales excepto las *valoraciones*. Tal y como indicamos en la sección 4, finalmente descartamos también los *destacados*, debido a su baja representatividad en la colección. A partir de ahí, consideramos tres familias de anotaciones: *etiquetas*, *comentarios* (conteniendo *notas* y *críticas*), y *contenido*.

3. Trabajo Relacionado

Existen varios trabajos en la literatura que analizan la utilidad de las anotaciones sociales para tareas de gestión de documentos web. Por lo que a la recuperación de información se refiere, varios trabajos han estudiado la inclusión de datos de marcadores sociales para modificar la búsqueda web (Hotho et al., 2006) (Yanbe et al., 2007) (Bao et al., 2007). En Heymann, Koutrika, y Garcia-Molina (2008) se sugiere que las URLs anotadas por usuarios en los marcadores sociales no alcanzan la escala de un buscador web, y por ello resulta difícil que puedan ayudar. No obstante, los autores consideran que puede resultar una opción muy interesante en el futuro si el uso de estos marcadores sociales sigue creciendo como hasta ahora.

Las anotaciones sociales se han utilizado también para tareas de organización de páginas web. En Ramage et al. (2009), la inclusión de información de etiquetas asignadas por usuarios mejora los resultados del clustering de páginas web basado en contenido textual. En cuanto a la clasificación de páginas web, Noll y Meinel (2008a) realizan un interesante estudio de las características de las anotaciones sociales. Indican que las etiquetas resultan más apropiadas para una clasificación de alto nivel, y también observan que los usuarios tienden a anotar las páginas principales de los sitios, mientras que las páginas más profundas quedan sin anotar. En un trabajo previo, Noll y Meinel (2007) concluyen que las etiquetas proveen información adicional no siempre contenida en el propio texto de la página.

En Noll y Meinel (2008b), estudian tres tipos de metadatos: anotaciones sociales (etiquetas), textos ancla -el texto de los enlaces entrantes-, y consultas de usuarios. En su análisis consideran las etiquetas como los metadatos más apropiados para resolver de qué tratan las páginas, y mejorar su clasificación.

A diferencia de los trabajos citados, en este estudio realizamos un análisis del uso de diferentes anotaciones sociales y el contenido textual para la clasificación de páginas web. Los utilizamos tanto de forma separada como combinándolos, con el fin de analizar el impacto de cada fuente de datos a la tarea de clasificación.

4. Colección de Datos

Hemos basado nuestra experimentación en una colección anotada de páginas web populares. Como entrada para la lista de URLs sobre la que generar la colección, nos basamos en el canal de actividad reciente de Delicious, limitándolo a aquellas páginas anotadas por al menos 100 usuarios. Una página anotada por al menos 100 usuarios asegura ser socialmente popular, además de proveer un conjunto de *etiquetas* que ha convergido (Golder y Huberman, 2006). Monitorizamos el canal de actividad reciente durante tres semanas entre diciembre de 2008 y enero de 2009. De esta manera, obtuvimos una lista de 87.096 URLs únicas.

Por otro lado, utilizamos el Open Directory Project[4] (ODP) como *Gold Standard* para clasificación de páginas web. Así, encontramos que 12.616 de nuestras URLs también se encontraban en ODP. En cuanto a la taxonomía, nos basamos en las 17 categorías que componen el primer nivel de la taxonomía de ODP. Como unas pocas páginas caen en más de una categoría, decidimos seleccionar una de ellas de forma aleatoria. Analizando la naturaleza de nuestra colección de datos, encontramos que la distribución de los documentos sobre las categorías no está balanceada, variando desde 39 hasta 3.289 documentos.

Una vez teníamos esta lista de URLs clasificadas, obtuvimos la siguiente información para cada URL:

Número de usuarios que lo han guardado en Delicious: no tiene por qué coincidir con el número de usuarios que lo ha anotado, ya que esta acción es opcional.

Lista Top 10 de *etiquetas* en Delicious: incluye las *etiquetas* más utilizadas junto con su número de ocurrencias.

[4] http://www.dmoz.org

Notas en Delicious: aunque el sistema limita el número de *notas* a 2.000, consideramos que es suficiente para nuestros objetivos.

Actividad completa de etiquetas (ACE) en Delicious: el sitio provee una lista de la actividad de usuarios para una URL, incluyendo las *etiquetas* que éstos han asignado. Esta lista está limitada a 2.000 usuarios.

Críticas en StumbleUpon: cabe destacar que 2.697 URLs de nuestra colección no tenían información de *críticas* en este sitio. Esto muestra que las *críticas* son insuficientes para clasificar por sí solas, pero pueden resultar interesantes combinándolas con otros datos.

Destacados en Diigo: sólo 1.920 URLs disponían de esta información, por lo que los consideramos insuficientes y decidimos no usarlos en el estudio.

Resumiendo, la colección se compone de 12.616 URLs únicas, clasificadas sobre el primer nivel de ODP y con sus correspondientes *etiquetas* y *notas* de Delicious, así como *críticas* de StumbleUpon.

5. *Máquinas de Vectores de Soporte*

En este trabajo, utilizamos las máquinas de vectores de soporte (SVM) (Joachims, 1998) para realizar las tareas de clasificación. Esta técnica utiliza el modelo espacio vectorial para la representación de documentos, y asume que los documentos de la misma categoría caerán en espacios separables. Sobre esto, busca un hiperplano que separe las clases; este hiperplano debe maximizar la distancia entre él y los documentos más cercanos, lo que se conoce como margen.

Aunque la aproximación SVM tradicional realiza sólo tareas binarias, se han propuesto diferentes adaptaciones a tareas multiclase (Weston y Watkins, 1999). Un clasificador SVM multiclase para k categorías define otros tantos hiperplanos en la fase de entrenamiento, con lo que cada uno de ellos separa los documentos correspondientes a esa clase del resto. En la fase de clasificación, al realizar predicciones para cada nuevo documento, el clasificador es capaz de establecer un margen sobre cada uno de los hiperplanos. Estos márgenes hacen referencia a la confianza que se tiene sobre si un documento pertenece a cada una de las clases. Como resultado, el clasificador ofrece como predicción aquella clase que maximiza el margen.

6. *Clasificación con Anotaciones Sociales*

Realizamos varios experimentos utilizando diferentes representaciones de los documentos. En estos experimentos, cambia la forma en la que se representan los documentos de forma vectorial, así como los datos que se utilizan como entrada, pero la configuración general se mantiene intacta.

Para llevar a cabo la clasificación multiclase con SVM, utilizamos SVMmulticlass (Joachims, 1999). Los experimentos se ejecutaron utilizando los parámetros por defecto recomendados por el propio autor, sobre un kernel polinomial. Aunque esta configuración pudiera optimizarse, no es el objetivo de este trabajo, ya que buscamos evaluar las diferencias de rendimiento usando diferentes representaciones. Como medida de evaluación nos basamos en el acierto, definido como el porcentaje de documentos correctamente clasificados.

Para cada una de las representaciones, se realizaron varias selecciones de conjuntos de entrenamiento y test. Por una parte, se establecieron diferentes tamaños para el conjunto de entrenamiento, desde 200 hasta 3.000 documentos. Esto permite evaluar la evolución de los resultados cuando los datos de entrenamiento aumentan. Por otro lado, se realizaron 6 ejecuciones para cada tamaño de entrenamiento. Entre estas ejecuciones varían los documentos seleccionados para entrenamiento, habiendo realizado una selección aleatoria. Finalmente calculamos la media para estas ejecuciones, que es la que mostramos como resultado, formando una línea con los diferentes tamaños de entrenamiento.

A continuación analizamos y evaluamos las diferentes propuestas de representación basadas en *etiquetas* y *comentarios*, de forma separada.

6.1. Clasificación con Etiquetas

Trabajos previos sugieren que las *etiquetas* pueden utilizarse para clasificación, pero ¿cuál es la mejor manera de usarlas? Evaluamos y comparamos las siguientes:

Etiquetas sin pesos: únicamente se tiene en cuenta si una *etiqueta* aparece o no en la lista top 10 de una página web. Esta aproximación ignora los pesos de cada *etiqueta*, asignando un simple valor binario.

Etiquetas ordenadas: se tiene en cuenta el puesto en el que se encuentra cada una las

etiquetas que forman el top 10. Se asigna un valor de 1 al que se encuentra en primer lugar, 0,9 al segundo, 0,8 al tercero, y así sucesivamente. Esta aproximación respeta la posición de cada *etiqueta*, pero ignora sus pesos.

Porcentaje de usuarios: además de la información facilitada por las *etiquetas* en el top 10, se tiene en cuenta el número de usuarios que han guardado la página. Así, se calcula el porcentaje de usuarios que ha anotado cada *etiqueta* del top 10. Una *etiqueta* habrá sido anotada por el 100 % de los usuarios si su peso coincide con el número de usuarios que guarda la página. Esta aproximación considera de alguna manera los pesos de cada *etiqueta*, pero se pierden los órdenes de magnitud al convertirlos en relaciones.

Etiquetas con pesos (Top 10): se tiene en cuenta el peso de cada *etiqueta* del top 10, pero se ignora el número de usuarios que guarda cada página. Nótese que se mezclan diferentes órdenes de magnitud en esta representación, al contener páginas que varían desde 100 hasta unas 61.000 anotaciones.

Etiquetas con pesos (ACE): al igual que la anterior, esta aproximación se basa en el peso de las *etiquetas*, aunque se tiene en cuenta la actividad completa de *etiquetas* en lugar de considerarse únicamente las del top 10. Para reducir la dimensionalidad de los vectores, relajando el coste computacional y manteniendo la representatividad, eliminamos todas las *etiquetas* que únicamente aparecen en una página de nuestra colección.

Hay que tener en cuenta que la actividad completa de las *etiquetas* se limita a la contribución de los últimos 2.000 usuarios, mientras que el top 10 tiene en cuenta todo el histórico, pudiendo participar más de 2.000 usuarios. En nuestra colección, existen 957 páginas web anotadas por más de 2.000 usuarios, con una media de 5.329 usuarios por página.

Las 4 primeras aproximaciones generan vectores dispersos de 12.116 dimensiones, donde sólo 10 de ellas tienen un valor diferente de 0 para cada página web; mientras que la última resulta en vectores de 175.728 dimensiones.

Los resultados para las aproximaciones basadas en *etiquetas* se muestran en la figura 1. Destaca la superioridad de las aproximaciones basadas en los pesos de las *etiquetas*, sobre todo cuando crece el conjunto de entrenamiento. Por otro lado, las representaciones basadas en el orden, utilizando fracciones y sin pesar, quedan muy por detrás. Parece claro que la consideración de los pesos de las *etiquetas* es una aproximación interesante que aporta información útil. Sólo es mejorada por la aproximación sin pesos cuando el conjunto de entrenamiento es de sólo 200 documentos. Cuando se aumenta el conjunto de entrenamiento, además, el rendimiento de las aproximaciones con pesos sigue mejorando, mientras que se mantiene constante para el resto, con un comportamiento asintótico.

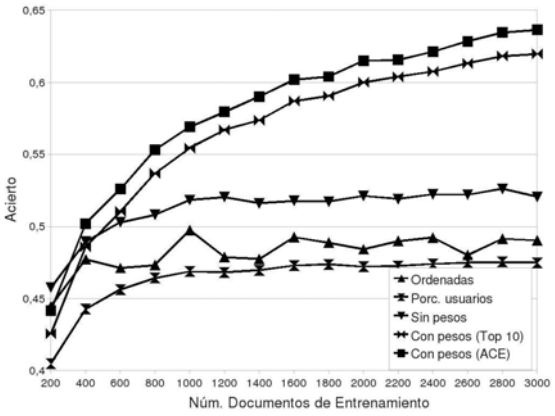

Figura 1: Resultados utilizando etiquetas

Nuestra conjetura es que el aprendizaje basado en sólo 200 documentos es mucho menos representativo para las aproximaciones con pesos, ya que sus valores son mucho más dispersos. Esto hace que la fase de entrenamiento sea menos completa que para el resto de aproximaciones, debido a que se basan en valores entre 0 y 1. De este modo, al aumentar el número de documentos de entrenamiento se invierte este comportamiento.

Al comparar las dos aproximaciones con pesos, vemos que cuantas más *etiquetas* se tienen en cuenta para la representación, mayor es el rendimiento del clasificador. La aproximación basada en la actividad completa de *etiquetas* supera siempre a la del top 10. La información adicional aportada por la primera hace mejorar los resultados, con una diferencia constante respecto a la segunda.

Concluimos con que los pesos de las *etiquetas* aportan información útil para clasificación temática, mejor incluso cuando se utiliza independientemente del número de usuarios que guardan la página. Por tanto, resulta más importante basarse en cómo han anotado los usuarios, y no tanto en cuántos usuarios

lo han hecho.

6.2. Clasificación con Comentarios

En lo que a los *comentarios* se refiere, disponemos de dos metadatos: *notas* y *críticas*. Ambos son textos que describen, definen o, de alguna manera, hacen referencia a una página web. Como hemos indicado, una cantidad significativa de documentos no dispone de *críticas* en nuestra colección, por lo que no son suficientes por sí solas para clasificar. Esta información podría resultar útil al combinarla con *notas*, ofreciendo información adicional. De esta manera, hemos experimentado las siguientes aproximaciones:

Sólo notas: se representa cada página con las *notas* escritas por los usuarios. En primer lugar, se unen todas las *notas* como si fueran sólo una. Una vez unidas, se obtiene la representación vectorial utilizando la función Term Frequency-Inverse Document Frequency (TF-IDF), y eliminando de la representación aquellos términos con una frecuencia de documentos (DF) baja. Estos vectores tienen 83.959 dimensiones.

Uniendo notas y críticas: se consideran también las *críticas*. Así como unimos todas las *notas* para una página, hacemos lo mismo con las *notas* y *críticas*. De la misma manera, la representación vectorial se basa en TF-IDF. Los vectores resultantes se componen de 95.149 dimensiones.

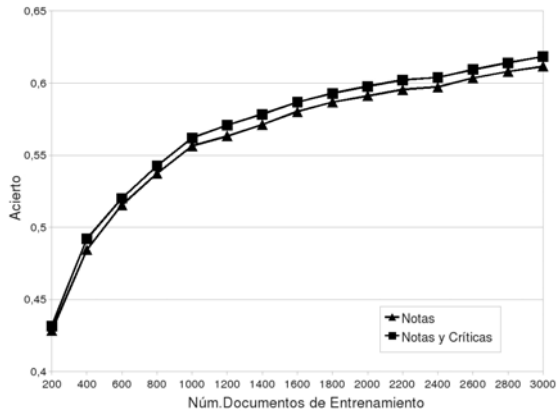

Figura 2: Resultados utilizando comentarios

La figura 2 muestra los resultados para las aproximaciones basadas en *comentarios*. Aunque ambas ofrecen unos resultados similares, la gráfica muestra una ligera superioridad para la aproximación que también considera *críticas*. Aunque las *críticas* son supuestamente subjetivas, y podrían parecer inicialmente dañinas para clasificación temática, estos resultados muestran que también pueden resultar útiles para esta tarea.

6.3. Contenido vs Anotaciones

Una vez probada la clasificación con *etiquetas* y con *comentarios*, y obtenidas las mejores aproximaciones para cada uno de estos metadatos, nuestro objetivo es el de compararlas con el *baseline*, la clasificación basada en *contenido*. Esta comparación incluye:

Contenido: consideramos el texto plano extraído tras eliminar las marcas html. Este *contenido* se representa utilizando la función TF-IDF, eliminando de la representación aquellos términos con una frecuencia de documentos (DF) baja, resultando en vectores de 156.028 dimensiones.

Comentarios: utilizamos directamente la mejor aproximación para *comentarios*, aquélla que combina *notas* y *críticas*.

Etiquetas: se utiliza la mejor aproximación, usando *etiquetas* con pesos sobre la actividad completa de *etiquetas*.

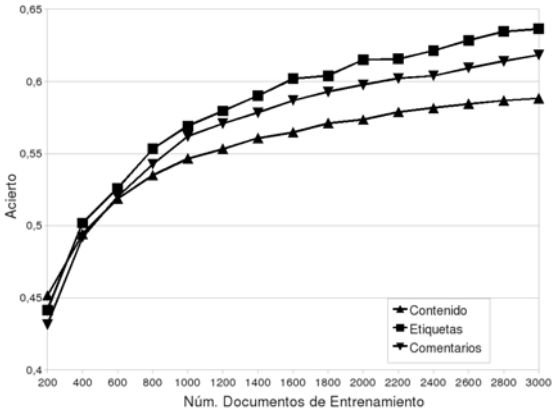

Figura 3: Resultados utilizando etiquetas, comentarios y contenido

La figura 3 muestra los resultados de esta comparativa de aproximaciones. Se puede ver que ambas anotaciones sociales mejoran el *baseline*, la clasificación basada en *contenido*. Una vez más, se da una excepción cuando sólo se entrena con 200 documentos, donde el *contenido* supera al resto. No obstante, según aumenta el tamaño del conjunto de entrenamiento, la aproximación basada en *contenido* se queda atrás.

Comparando el comportamiento de las anotaciones sociales, se obtiene un rendi-

miento superior para la aproximación basada en *etiquetas*. Analizando el caso con 3.000 documentos de entrenamiento, las *etiquetas* superan en un 8 % al *contenido*, y en un 3 % a los *comentarios*.

Entrando en más detalles, se pueden analizar los documentos mal clasificados. SVM devuelve una serie de márgenes para cada documento, un margen por cada clase, por lo que se puede extraer una lista de posiciones de las predicciones. Para nuestros experimentos sobre 17 categorías, la mejor posición para la categoría correcta en un documento mal clasificado sería segundo, mientras que el peor sería decimoséptimo. La figura 4 muestra el número de documentos que cae en cada posición de ese ranking para un conjunto de entrenamiento de 3.000 documentos.

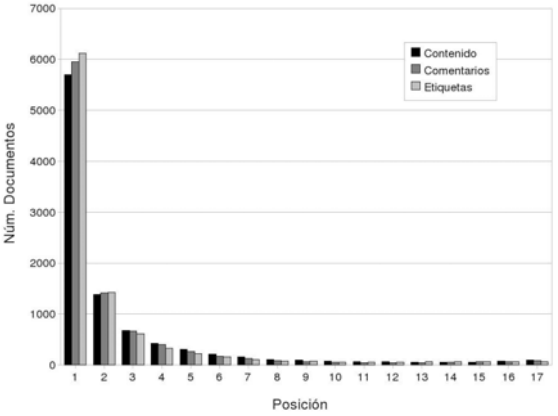

Figura 4: Número de predicciones correctas en cada posición del ranking

De la misma manera, la figura 5 muestra la posición media en la que cae la categoría correcta según las predicciones de cada clasificador. Estos gráficos muestran, nuevamente, la superioridad de las *etiquetas* respecto a los *comentarios* y el *contenido*. Además, como muchos de los documentos mal clasificados tienen una posición alta (segundo o tercero), y sólo unos pocos caen en la cola, los clasificadores muestran una alta fiabilidad.

7. Combinación de Clasificadores

Aunque la representación basada en *etiquetas* supera a las otras dos, todas ellas muestran resultados alentadores, y parecen lo suficientemente buenas como para combinarlas mejorando los resultados. Una interesante propuesta para su combinación es la de los clasificadores unificados (Sun et al., 2004). Los

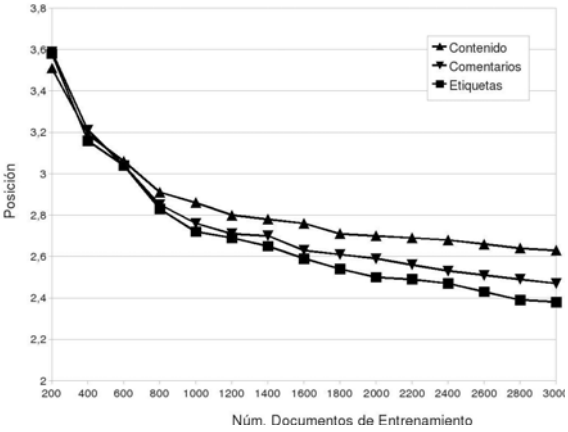

Figura 5: Posición media de las predicciones correctas

clasificadores unificados se basan en las predicciones de cada uno de los clasificadores, y se combinan utilizando una función de decisión, la cual sirve para definir un peso o una relevancia a cada clasificador.

La combinación de clasificadores SVM se realiza comúnmente por medio de la suma de sus márgenes (o fiabilidades) para cada clase. Cada documento tendrá por tanto una nueva suma para cada clase. La clase que maximice esta suma será la que prediga el sistema final.

En este estudio, experimentamos todas las posibilidades con las mejores aproximaciones para *etiquetas*, *comentarios* y *contenido*: (a) *etiquetas* + *contenido*; (b) *etiquetas* + *comentarios*; (c) *comentarios* + *contenido*, y (d) *etiquetas* + *comentarios* + *contenido*.

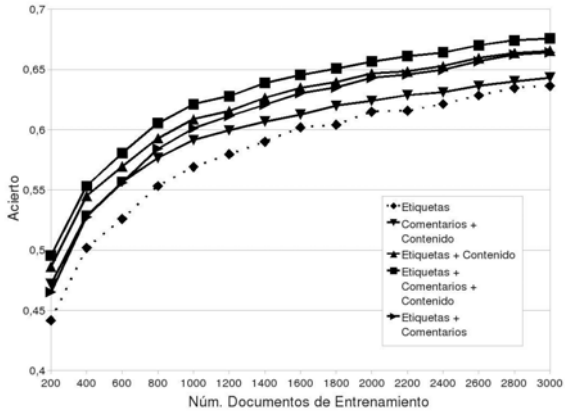

Figura 6: Resultados de la combinación de clasificadores

Los resultados para los clasificadores unificados se muestran en la figura 6. Se incluye la

línea correspondiente a la clasificación basada en *etiquetas*, como referencia para comparar con la mejor aproximación no combinada.

Cuando se combinan diferentes clasificadores, algunos de los errores de unos son corregidos por otros, como muestran los resultados; cualquiera de las combinaciones supera a la clasificación basada en *etiquetas*.

Entre las combinaciones, el mejor resultado es para la triple combinación de metadatos, superando a cualquiera de las combinaciones dobles. De media, la triple combinación es un 2 % superior que la mejor de las combinaciones dobles. Para las combinaciones dobles, el rendimiento es mayor cuando se tienen en cuenta las *etiquetas*, mostrándose nuevamente que son los mejores metadatos para este propósito.

La combinación de *etiquetas* y *contenido* supera a la de *etiquetas* y *comentarios*, mientras que esta última supera a la de *contenido* y *comentarios*. De ello se deduce que las *etiquetas* son las anotaciones que más aportan, seguidas del *contenido* y los *comentarios*.

Analizamos también la contribución de cada clasificador en el resultado final. La figura 7 muestra el acierto desglosado por clasificadores. Se puede ver que la principal diferencia en el rendimiento final depende del número de documentos correctamente clasificados por los 3 clasificadores. Mientras que el número de aciertos permanece estable para los casos donde 2 o menos clasificadores aciertan, los casos de acierto en los que los 3 clasificadores habían acertado aumenta según crece el conjunto de entrenamiento. Obviamente, la predicción final será siempre la correcta cuando los 3 clasificadores han acertado, lo que ocurre más frecuentemente para conjuntos de entrenamiento de mayor tamaño.

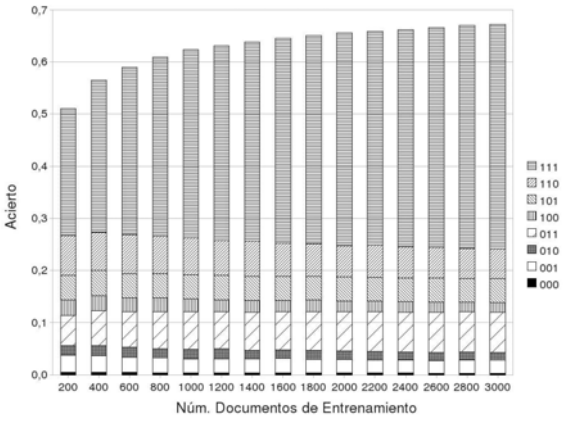

Figura 7: Aportación de cada clasificador a la combinación triple
(La leyenda muestra tres dígitos binarios para los clasificadores basados en contenido, comentarios y etiquetas, respectivamente; un 1 es un acierto, mientras un 0 se refiere a un error)

8. Conclusiones y Trabajo Futuro

En este trabajo, hemos analizado y evaluado la utilización de anotaciones sociales para clasificación de páginas web sobre la taxonomía de Open Directory Project. Aunque los marcadores sociales proveen diversos tipos de anotaciones, algunos de ellos no son lo suficientemente populares aún como para poder considerarlos. Hemos observado que las *etiquetas* y *comentarios* tienen suficiente cobertura para su aplicación a la clasificación de páginas web. Nuestros experimentos de clasificación con *etiquetas* y *comentarios* muestran resultados alentadores, que mejoran la clasificación con *contenido textual*. Cuanto mayor es el conjunto de entrenamiento utilizado, mayor es la ventaja a favor de las *etiquetas* y los *comentarios* sobre el *contenido*. Además, hemos demostrado que la combinación de anotaciones sociales con el propio *contenido* mejora aún más el rendimiento.

Nuestros experimentos corroboran las conclusiones de Noll y Meinel (2007) y Noll y Meinel (2008a) en lo que respecta a la utilidad de las *etiquetas* para clasificación; particularmente, al realizar una clasificación de alto nivel, como la utilizada en este estudio.

Como trabajo futuro, queda por evaluar la utilidad de las anotaciones sociales para una clasificación de más bajo nivel. La utilización de *destacados* como entrada del clasificador es otro aspecto a tener en cuenta, si en algún momento llega a alcanzar un mayor grado de popularidad. Por otro lado, podría realizarse un estudio más profundo: (a) de las *etiquetas*, tratando de filtrar aquéllas que pudieran considerarse subjetivas, o tratando de resolver los problemas de sinonimia y polisema que presentan, y (b) de los *comentarios*, filtrando aquéllos que pueden generar ruido.

Agradecimientos

Trabajo subvencionado parcialmente por la red de investigación MAVIR (S-0505/TIC-0267), la Consejería de Educación de la Comunidad de Madrid y el proyecto QEAVis-Catiex (TIN2007-67581-C02-01) del Ministerio de Ciencia e Innovación.

Bibliografía

Bao, Shenghua, Guirong Xue, Xiaoyuan Wu, Yong Yu, Ben Fei, y Zhong Su. 2007. Optimizing web search using social annotations. En *Proceedings of the 16th international conference on World Wide Web*, páginas 501–510, Banff, Alberta, Canada. ACM.

Golder, Scott y Bernardo A. Huberman. 2006. The structure of collaborative tagging systems. *Journal of Information Science, 32(2)*, páginas 198–208.

Heymann, Paul, Georgia Koutrika, y Hector Garcia-Molina. 2008. Can social bookmarking improve web search? En *Proceedings of the international conference on Web search and web data mining*, páginas 195–206, Palo Alto, California, USA. ACM.

Hotho, Andreas, Robert Jäschke, Christoph Schmitz, y Gerd Stumme. 2006. Information retrieval in folksonomies: Search and ranking. En *The Semantic Web: Research and Applications*, páginas 411–426.

Joachims, Thorsten. 1998. Text categorization with support vector machines: Learning with many relevant features. En *European Conference on Machine Learning (ECML)*, páginas 137–142, Berlin. Springer.

Joachims, Thorsten, 1999. *Making large-scale support vector machine learning practical*, páginas 169–184. MIT Press.

Noll, Michael G. y Christoph Meinel. 2007. Authors vs. readers: a comparative study of document metadata and content in the www. En *Proceedings of the 2007 ACM symposium on Document engineering*, páginas 177–186, Winnipeg, Manitoba, Canada. ACM.

Noll, Michael G. y Christoph Meinel. 2008a. Exploring social annotations for web document classification. En *Proceedings of the 2008 ACM symposium on Applied computing*, páginas 2315–2320, Fortaleza, Ceara, Brazil. ACM.

Noll, Michael G. y Christoph Meinel. 2008b. The metadata triumvirate: Social annotations, anchor texts and search queries. En *Web Intelligence and Intelligent Agent Technology, 2008. WI-IAT '08. IEEE/WIC/ACM International Conference on*, volumen 1, páginas 640–647.

Ramage, Daniel, Paul Heymann, Christopher D. Manning, y Hector Garcia-Molina. 2009. Clustering the tagged web. En *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, páginas 54–63, Barcelona, Spain. ACM.

Sun, Bing-Yu, De-Shuang Huang, Lin Guo, y Zhong-Qiu Zhao, 2004. *Support Vector Machine Committee for Classification*, páginas 648–653.

Weston, J. y C. Watkins. 1999. Multi-class support vector machines. En *Proceedings of ESAAN '99, the European Symposium on Artificial Neural Networks*.

Yanbe, Yusuke, Adam Jatowt, Satoshi Nakamura, y Katsumi Tanaka. 2007. Can social bookmarking enhance search in the web? En *Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries*, páginas 107–116, Vancouver, BC, Canada. ACM.