Categorización de texto sensible al coste para el filtrado de contenidos inapropiados en Internet*

José María Gómez Hidalgo
Enrique Puertas Sánz
Francisco Carrero García
Manuel de Buenaga Rodríguez
Departamento de Inteligencia Artificial
Universidad Europea de Madrid
Villaviciosa de Odón, 28670, Madrid (Spain)
{jmgomez,epuertas,fcarrero,buenaga}@.uem.es

Resumen: El creciente problema del acceso a contenidos inapropiados de Internet se puede abordar como un problema de categorización automática de texto sensible al coste. En este artículo presentamos la evaluación comparativa de un rango representativo de algoritmos de aprendizaje y métodos de sensibilización al coste, sobre dos colecciones de páginas Web en español e inglés. Los resultados de nuestros experimentos son prometedores.
Palabras clave: Categorización automática de texto, filtrado de Internet, aprendizaje sensible al coste, Receiver Operating Characteristic.

Abstract: The access to inapropiate Internet content is an increasing problem that can be approached as a cost-sensitive Automated Text Categorization task. In this paper, we report a series of experiments that compare a representative range of learning algorithms and methods for making them cost-sensitive, on two Web pages collections in Spanish and English. The results of our experiments are promising.
Keywords: Automated Text Categorization, Internet Filtering, Cost Sensitive Learning, Receiver Operating Characteristic.

1. Introducción

Es indudable que Internet y su implantación progresiva en todos los ambientes de nuestra sociedad (la llamada Sociedad de la Información) acarrea notables beneficios para sus usuarios, posibilitando nuevas formas de comunicación, trabajo y educación.

Sin embargo, la naturaleza inherentemente distribuida y de difícil control de Internet conlleva asimismo riesgos importantes para su aprovechamiento. En concreto, uno de los problemas más notables en la actualidad es el acceso a contenidos inapropiados por parte de niños y jóvenes, y por parte de los profesionales en su entorno de trabajo. Por una parte, existe la posibilidad de que, tanto activamente como pasivamente, los niños y jóvenes accedan a contenidos de Internet que son incapaces de juzgar correctamente, incluyendo los pornográficos, o los que promueven la violencia, el racismo y la adhesión a sectas. Por la otra, los trabajadores acceden a contenidos similares o de otro tipo (información sobre búsqueda de empleo, contenidos de entretenimiento, y otros) en su entorno laboral, haciendo uso de los recursos de la empresa con un fin para el que no están destinados, e incurriendo en abuso (ACM, 2002).

En ambos casos, y ante la falta de regulaciones apropiadas, es adecuada la adopción de sistemas de filtrado y monitorización para limitar el acceso a contenidos inapropiados de Internet. Nuestra institución interviene en dos proyectos de I+D llamados POESIA y TEFILA, orientados al desarrollo de herramientas para el filtrado de contenidos inapropiados en Internet en ambos entornos. En concreto, el objetivo de POESIA (Public Opensource Environment for a Safer Internet Access) es el desarrollo de un sistema de código abierto para el filtrado de información inapropiada en ambientes escolares, por medio de la integración de técnicas avanzadas de Ingeniería del Lenguage Natural, Aprendizaje Automático, procesamiento de imágenes, y otras. Por otro lado, en TEFILA (TEcnicas de Filtrado basadas en Ingeniería del Len-

* Esta investigación ha sido financiada parcialmente por la Comisión Europea a través del Safe Internet Action Plan (POESIA - SIAP-2117) y por el Ministerio de Ciencia y Tecnología a través del programa PROFIT (FIT-070000-2002-861).

guage, Aprendizaje automático y agentes) se persigue el desarrollo de técnicas más efectivas, flexibles y configurables que las actuales para la construcción de herramientas de filtrado en el puesto de trabajo.

En nuestro trabajo, abordamos la detección de contenidos inapropiados como un problema de categorización automática de texto (Automated Text Categorization, ATC) (Sebastiani, 2002). Esta tarea consiste en la clasificación automática de documentos en categorías predefinidas. En nuestro caso, y para la investigación presentada en este trabajo, los documentos son páginas Web, y las categorías son PORNOGRAFÍA y SEGURO, en referencia al tipo de información más comunmente filtrada por las herramientas actuales.

La construcción de sistemas de categorización automática puede realizarse de manera manual (derivando conjuntos de reglas de clasificación, al estilo de los sistemas expertos) o parcialmente automática (usando técnicas de Recuperación de Información y de Aprendizaje Automático para construir sistemas de clasificación a partir de ejemplos manualmente clasificados). Este segundo enfoque, llamado en la literatura *basado en aprendizaje* (Sebastiani, 2002), involucra usualmente la representación de los documentos (páginas Web) como vectores de pesos de términos, sobre los que se aplican algoritmos de aprendizaje que extraen modelos de clasificación llamados clasificadores. En la actualidad, los clasificadores obtenidos por estos medios son casi tan precisos como los seres humanos, especialmente en la categorización temática (usando sistemas de categorías orientados al tema de un documento, como los bibliográficos o los de los directorios Web como Yahoo!).

Sin embargo, los sistemas usuales de ATC basados en aprendizaje usualmente no tienen en cuenta que los distintos tipos de errores que puede cometer el sistema tiene costes distintos para el usuario. En nuestro caso, es más perjudicial que un niño acceda a un contenido pornográfico (un error por defecto) que el que no pueda acceder a un contenido válido (error por exceso). Para resolver este problema, es necesario aplicar métodos de aprendizaje sensibles al coste de los errores, que son capaces de tener en cuenta estos costes asimétricos para derivar clasificadores que prefieren unos tipos de errores sobre otros.

En este trabajo presentamos unas series de experimentos orientadas a evaluar la efectividad de una gama representativa de algoritmos de aprendizaje (incluyendo los clasificadores bayesianos ingenuos, la inducción de árboles de decisión con C4.5, la generación de reglas de clasificación y las Support Vector Machines) adaptados al coste por medio de distintas estrategias (basada en umbral, basada en pesos y MetaCost), para la detección de contenidos pornográficos en el WWW, en inglés y castellano. Los distintos enfoques han sido evaluados utilizando el método Receiver Operating Characteristic Convex Hull, que resulta el más adecuado cuando los costes de los errores son asimétricos pero desconocidos en condiciones reales.

El resto de este trabajo está organizado de la siguiente manera. En primer lugar, presentamos el modelo de categorización basado en aprendizaje sensible al coste. En segundo lugar, describimos en método de evaluación usado en nuestros experimentos. A continuación, describimos el entorno experimental, detallando los enfoques evaluados y la colección de evaluación. Después se presentan y analizan los resultados de nuestros experimentos, para finalmente extraer las conclusiones y esbozar las lineas de trabajo futuro.

2. *Categorización automática sensible al coste*

La ATC basada en aprendizaje es hoy en día un modelo de categorización sólido y efectivo. Usualmente, este modelo no tiene en cuenta que la clasificación de contenidos inapropiados es un problema con costes asimétricos, es decir, en el unos tipos de errores son más dañinos que otros. En este apartado presentamos el modelo general de la ATC basada en aprendizaje, y la adaptación del mismo a entornos en los que los costes de los errores son asimétricos.

2.1. Categorización automática basada en aprendizaje

La construcción de sistemas de ATC basada en aprendizaje se basa en diversos elementos tomados de los campos de la Recuperación de Información y del Aprendizaje Automático (Sebastiani, 2002). Esencialmente, el proceso consiste en la representación de un conjunto de documentos (en nuestro caso, páginas Web – documentos en formato HTML) manualmente clasificados (llamado *colección de entrenamiento*) como vectores de pesos de términos, a los que se aplica

un algoritmo de aprendizaje que construye un modelo de clasificación o clasificador. Los documentos a clasificar se representan de igual forma, de modo que el clasificador es capáz de asignar una categoría (en nuestro caso, PORNOGRAFÍA o SEGURO) a los mismos. Este modelo está basado en diversos elementos:

1. El método de representación o *indexación* de los documentos. El más usual consiste en representar los documentos como vectores de pesos de términos, según el Modelo del Espacio Vectorial para la Recuperación de Información (Salton, 1989). En este modelo, llamado en la literatura "bag of words", los términos son normalmente palabras aisladas a las que se aplica un extractor de raíces como el de Porter, una vez que se han eliminado aquellas más frecuentes o con menor carga semántica usando una lista de parada. Los pesos de los términos en cada documento se puede definir de varias maneras, incluyendo los pesos binarios (1 si el término aparece en el documento, y 0 en caso contrario), los pesos TF (Term Frequency o frecuencia del término en el documento), o los pesos TF.IDF (el anterior multiplicado por la Inverse Document Frequency o frecuencia inversa en documentos, definida usualmente como $\log_2(n/df(t))$, donde n es el número de documentos de entrenamiento, y $df(t)$ el número de documentos en los que aparece el término t).
2. El método de selección de términos. Con el fin de evitar el sobre ajuste en el aprendizaje, y para aumentar su eficiencia y efectividad, se suele seleccionar un subconjunto de términos de los originales. Para ello, se utiliza una métrica de calidad de los términos, y se seleccionan aquellos cuyo valor para la misma es alto. En (Yang y Pedersen, 1997) se demuestra experimentalmente que la utilización de las métricas Ganancia de Información (Information Gain, IG) y χ^2 permite eliminar hasta un 99 % de los términos originales, lográndose un importante aumento de la eficiencia e incluso un ligero aumento de la efectividad.
3. El algoritmo de aprendizaje. Son múltiples los algoritmos de aprendizaje aplicados a problemas de ATC en la literatura, incluyendo los clasificadores probabilísticos como el bayesiano ingenuo, la inducción de árboles de decisión con C4.5, la generación de reglas de clasificación con Ripper, las Support Vector Machines – SVM, y otros (ver la recopilación de (Sebastiani, 2002), y la comparativa de (Yang, 1999)). La efectividad de los algoritmos es variable, siendo las SVM uno de los más efectivos.

Este modelo es muy efectivo en situaciones en las que la categorización es temática, es decir, se pretende asignar uno o varios temas a un documento en función de su contenido. En (Sebastiani, 2002) se argumenta que, hoy por hoy, la ATC basada en aprendizaje es capaz de alcanzar grados de precisión similares a los del ser humano.

2.2. Aprendizaje sensible al coste para la categorización de texto

La mayoría de algoritmos de aprendizaje anteriores, por su propia naturaleza, buscan minimizar el número de errores del clasificador generado. Sin embargo, son múltiples los problemas de Aprendizaje Automático (y de ATC) en los que los errores cometidos por el clasificador generado no tienen la misma importancia (Provost y Fawcett, 2001).

En concreto, en la ATC orientada al filtrado de correo masivo no solicitado, o *spam*, es preferible que el sistema clasifique un mensaje no solicitado como legítimo antes que lo contrario (Gómez, 2002). Ello se debe a que es probable que el usuario del sistema elimine los no solicitados sin un examen excesivamente detallado, corriendo el riesgo de eliminar mensajes legítimos e importantes. En términos de costes, y asumiendo que la clase positiva (a detectar) es el correo masivo no solicitado, se dice que un error de tipo Falso Positivo (clasificar un mensaje legítimo como masivo no solicitado) tiene mayor coste que un error de tipo Falso Negativo (el opuesto).

Esta situación se produce también en la clasificación de páginas Web pornográficas. En entornos escolares, y debido a las posibles consecuencias sobre los niños y jóvenes, es preferible errar por exceso (clasificando páginas seguras como pornográficas) que por defecto (lo contrario). Por tanto, es necesario aplicar algoritmos de aprendizaje que sean sensibles al coste de los errores, y que al construir el clasificador prioricen evitar unos tipos

de errores sobre otros.

Los métodos de aprendizaje sensibles al coste suelen ser adaptaciones de algoritmos existentes, como los árboles de decisión (Ting, 1998) y otros. Sin embargo, existen estrategias que son independientes del algoritmo de aprendizaje utilizado. Estas estrategias, corrientemente denominadas de metaesquemas de aprendizaje, toman como entrada un algoritmo de aprendizaje, una colección de entrenamiento y una distribución de costes, y generan un clasificador basado en el algoritmo de aprendizaje y adaptado a los costes de los errores. Ejemplos de estas estrategias son:

- La basada en umbral (Thresholding – (Witten y Frank, 1999)), que se puede aplicar a todo algoritmo cuya salida sea un clasificador que emite valores numéricos (como probabilidades, similitudes, etc.). La idea es simple: si, por ejemplo, un clasificador $\Phi(d)$ asigna la clase positiva a un documento d a partir de un umbral ν (es decir, cuando $\Phi(d) > \nu$), el umbral se ajusta para que el clasificador sea más o menos conservador, usando para ello una subcolección de documentos de entrenamiento reservados para este fin.
- La basada en pesos de ejemplares (Instance Weighting – (Ting, 1998)), aplicable a cualquier algoritmo de aprendizaje. Esta basada en dar más peso a los documentos o ejemplares de una clase (e.g. PORNOGRAFÍA), a fin de que el algoritmo se concentre especialmente en clasificar correctamente estos ejemplares, minimizando el error sobre ellos. El peso asignado es proporcional al coste de los errores sobre dichos documentos.
- MetaCost (Domingos, 1999), aplicable a cualquier algoritmo de aprendizaje. Esta sofisticada técnica consiste en reetiquetar la colección de entrenamiento de acuerdo con la salida de un comité de clasificadores generados por el algoritmo base usando el método de *bagging*, y entrenar luego un clasificador sobre la colección reetiquetada.

La aplicación de estos métodos permite adaptar al coste los algoritmos empleados tradicionalmente en la ATC, generando clasificadores más efectivos en situaciones en las que la distribución de costes en los errores es asimétrica.

3. Evaluación de la categorización sensible al coste

La evaluación de los sistemas de ATC se basa usualmente en dos elementos principales: la disposición de una colección de evaluación, y la utilización de métricas de efectividad. Las métricas contabilizan el índice de aciertos y errores de un clasificador sobre la colección de evaluación.

3.1. Colecciones de evaluación

Una colección de evaluación es un conjunto de documentos manualmente clasificados sobre los que se evalúa un sistema de ATC. El ejemplo arquetípico de colección de evaluación para la ATC es la colección Reuters-21578, que contiene noticias en inglés clasificadas de acuerdo con categorías temáticas de carácter económico (indicadores económicos, monedas, bienes, etc.) (Sebastiani, 2002).

Las colecciones de evaluación suelen dividirse en dos partes: un fragmento de la colección se reserva para el entrenamiento, y otro fragmento se emplea en la evaluación. En Reuters se han realizado hasta 4 particiones, usadas en distintos trabajos.

Una alternativa a este enfoque, frecuente en Aprendizaje Automático, es la validación cruzada (k-fold cross validation). Dado un número natural k (frecuentemente 10), la colección se divide en k partes, de modo que cada parte mantiene la misma distribución de documentos en clases. Se realizan k pruebas con $k-1$ partes de entrenamiento y 1 de evaluación, y se promedian los resultados. A falta de una partición sólida de la colección de evaluación, éste es el enfoque más razonable.

3.2. Métricas de efectividad

Las métricas de efectividad más populares en la evaluación de sistemas de ATC proviene del campo de la Recuperación de Información, e incluyen la tasa de recuperación o *recall*, la precisión y la medida F_1, que combina ambas (Sebastiani, 2002).

Estas métricas no son adecuadas en situaciones en que los costes son asimétricos. Medidas como la exactitud con pesos (Weighted Accuracy) o el coste (asumiendo 0 para los aciertos, y el coste para cada error) son más adecuadas en situaciones en que los costes asociados a los errores son conocidos.

Sin embargo, los costes reales raramente son conocidos en el mundo real, y además éstos pueden cambiar de unos entornos a otros. Por ejemplo, en una escuela se puede priorizar que se bloquee el acceso a páginas pornográficas a riesgo de bloquear páginas seguras, mientras que en una empresa dedicada a la medicina puede preferir lo contrario.

En situaciones de costes reales desconocidos, como la que se da en el filtrado de pornografía en Internet y en la detección de correo masivo no solicitado, resulta más adecuado realizar la evaluación usando el método Receiver Operating Characteristic Convex Hull – ROCCH) (Provost y Fawcett, 2001), que permite comparar enfoques cuando los costes son desconocidos pero relevantes, y seleccionar el método más adecuado una vez que estos se fijan. Describimos este método a continuación[1].

3.3. El método ROCCH

El método ROCCH ya ha sido utilizado en ATC para la evaluación de sistemas de detección de correo masivo no solicitado (Gómez, 2002). Este método parte de construir gráficas tipo Receiver Operating Characteristic (ROC) para los clasificadores evaluados. Un gráfica ROC es similar a una gráfica recall-precisión, en la que se representan en el eje de abscisas la tasa de falsos positivos (False Positive Rate – FPR, definida como el porcentaje de ejemplares clasificados en la clase positiva perteneciendo a la clase negativa), y en el de ordenadas la de positivos reales (True Positive Rate – TPR, definida como el porcentaje de ejemplares clasificados en la clase positiva de los pertenecientes a la misma).

Para cada clasificador sensible al coste, y cada distribución de costes, se puede obtener un punto (FPR, TPR) que se representa en una gráfica ROC. Cuanto más cerca de la esquina superior izquierda se halla el punto, mejor es el clasificador. Se puede obtener una curva uniendo los puntos obtenidos para distintas distribuciones de costes, asumiendo interpolación lineal. La comparación entre dos gráficas se realiza de igual modo que en las gráficas recall-precisión.

Por resultados teóricos, se pueden descartar los puntos de un diagrama ROC que no se encuentran en el recubrimiento convexo superior (upper Convex Hull) del conjunto de puntos representados. Asimismo, dada una distribución de clases y de costes de cada tipo de error (que corresponda a una situación real concreta), es posible seleccionar el clasificador o clasificadores más eficaces en dichas condiciones. Ello se debe a que a cada conjunto de condiciones de distribuciones de clases y costes se corresponde con la pendiente de una recta que se puede desplazar de abajo a arriba, para encontrar el punto superior que maximiza la efectividad, o de otro modo, minimiza el coste.

Operativamente, el método ROCCH consta de los siguientes pasos:

1. Para cada método de clasificación sensible al coste, obtener una curva ROC del modo siguiente: (a) entrenar y evaluar el clasificador para un conjunto de distribuciones de costes representativas, obteniendo una serie de puntos (FPR,TPR)[2]; y (b) obtener el recubrimiento convexo superior de la serie de puntos, descartar los que no forman parte de ellos, y unir los siguientes por medio de rectas (interpolación lineal).
2. Obtener el recubrimiento convexo superior de todas las curvas representadas. Típicamente, unos algoritmos superarán a otros en un rango de abscisas, pues es difícil que un enfoque sea absolutamente superior a los demás.
3. Hallar el rango de pendientes para las cuales cada curva coincide con el recubrimiento convexo. De este modo, se obtiene un cuadro que define bajo que condiciones es mejor cada clasificador.
4. En caso de conocer la distribución de clases y costes del entorno operativo real, obtener la pendiente asociada a dichas distribuciones y seleccionar el mejor clasificador disponible según este análisis.

El método ROCCH permite la comparación visual de la efectividad de un conjunto de clasificadores, de manera independiente de la distribución de clases y costes(Provost y Fawcett, 2001). De este modo, la decisión de

[1] Por razones de espacio, hemos de ser breves en la descripción del método ROCCH. Para conocerlo en más profundidad, véase (Provost y Fawcett, 2001).

[2] Cada punto se obtiene evaluando sobre una colección de evaluación. Para que los resultados sean más fiables, y ante la falta de una partición estable de la colección de evaluación, es recomendable obtener cada punto por validación cruzada.

cual es el mejor método se puede retrasar hasta que las condiciones operativas reales sean conocidas, obteniendo al mismo tiempo información valiosa.

4. Diseño experimental

En esta sección describimos en detalle los experimentos realizados, centrándonos en la construcción de la colección de entrenamiento y evaluación, y en los enfoques evaluados.

4.1. La colección de evaluación

Hemos tomado como base el directorio Open Directory Project (ODP)[3]. Un directorio WWW es un conjunto de recursos organizados en categorías ordenadas jerárquicamente. El directorio WWW más popular en la actualidad es el incluído en el portal Yahoo!. Sin embargo, el ODP es el directorio de mayor tamaño de los existentes actualmente. En el ODP figuran más de 3,8 millones de recursos (páginas WWW, grupos de noticias, servidores FTP, etc) organizados en más de 460.000 categorías, y mantenido por casi 56.200 editores humanos de manera gratuita. A diferencia de Yahoo! y otros, este directorio es libre en varios sentidos, pero especialmente en que es posible descargar la totalidad de sus contenidos sin restricciones.

El ODP puede ser usado para la elaboración de colecciones de páginas WWW pornográficas y válidas en múltiples idiomas, pues sus contenidos están separados regionalmente, y organizados en secciones para adultos y para el público general. En concreto, el ODP contiene aproximadamente 2,5 millones de referencias en inglés válidas y 100.000 para adultos. En las secciones sobre España, el ODP contiene aproximadamente 100.000 referencias válidas y 1.000 para adultos. Utilizando un procesador y un robot software programados al efecto, hemos construido una colección de referencias del siguiente modo:

- Hemos recolectado todas las direcciones válidas y para adultos en inglés y en español, y hemos seleccionado de manera aleatoria un subconjunto de ellas, obteniendo 5.335 direcciones válidas y 1.021 direcciones para adultos en español, y 5.091 direcciones válidas y 1.002 direcciones para adultos en inglés.
- Hemos descargado todas las páginas accesibles en las direcciones anteriores, excluyendo aquellas que tardaban más de 10 segundos en responder, obteniendo 4.956 documentos válidos y 966 para adultos en español, y 2.570 documentos válidos y 129 para adultos en inglés.

En nuestra experiencia, estos datos pueden ser suficientes para realizar un entrenamiento bastante efectivo en la detección de contenidos pornográficos en inglés y español.

4.2. Procesamiento de la colección

Los documentos se han representado usando el Modelo del Espacio Vectorial (Salton, 1989), como vectores de pesos binarios de términos. Los términos se definen a partir de las palabras o secuencias de caracteres alfanuméricos, separadas por blancos u otros separadores, una vez eliminadas las etiquetas HTML de las páginas. Las palabras se filtran usando una lista de palabras vacía distinta para cada idioma, y se extrae su raíz usando el algoritmo de Porter. Se obtienen aproximadamente 16.600 términos distintos para el español y 11.900 para el inglés. A continuación, hemos seleccionado para cada idioma un 1 % de los términos originales, usado la Ganancia de Información como métrica de calidad de atributos. De este modo, se usan 166 términos en la representación de los documentos en español, y 119 para el inglés. Las clases SEGURO y PORNOGRAFÍA se consideran como clases positiva y negativa, respectivamente.

4.3. Enfoques evaluados

En este trabajo, hemos evaluado los siguientes algoritmos de aprendizaje:

- Bayes ingenuo (Lewis, 1998), que genera un modelo probabilístico de las categorías, asumiendo que las apariciones de los términos son independientes entre si.
- C4.5 (Quinlan, 1993), que induce un árbol de decisión en el que las ramas van etiquetadas con comprobaciones sobre los valores (pesos) de los atributos (términos), y cuyas hojas van etiquetadas con la clase mayoritaria de los ejemplares de entrenamiento que cumplen las comprobaciones que definen un camino hacia la hoja.
- PART (Frank y Witten, 1998), que produce listas de reglas de decisión. Cada regla consta de una conjunción de comprobaciones como antecedente, y una clase

[3] `http://dmoz.org`.

como consecuente. Las reglas se aplican consecutivamente, siendo la última una por defecto que asigna la clase mayoritaria de los ejemplares de entrenamiento no cubiertos por las demás reglas.

- SVM (Joachims, 2001), que genera en nuestra configuración una función lineal sobre los pesos de los términos, cuya aplicación sobre un nuevo documento da como resultado un valor numérico. Si es valor es mayor que cero, el documento se clasifica en la clase positiva, y en caso contrario en la negativa.

Cada uno de los algoritmos se ha hecho sensible al coste usando los tres meta-esquemas descritos en el apartado 2.2[4]. Para cada algoritmo, meta-esquema, e idioma, se han obtenido 41 puntos (FPR,TPR), por validación cruzada con $k = 10$. Estos puntos se corresponden con las relaciones de coste entre la clase positiva y negativa siguientes: 1/1000, 1/900, ..., 1/100, 1/90, ..., 1/10, 1/5, 1, 5, 10, 20, ..., 90, 100, ..., 900 y 1000.

5. *Resultados y análisis*

En la figura 1 presentamos el recubrimiento convexo de las curvas obtenidas para cada clasificador y meta-esquema de sensibilización al coste, en forma de dos curvas (del español - SP y del inglés – EN). Por claridad, omitimos las 20 curvas correspondientes a cada combinación de algoritmos e idiomas.

En la tabla 1 se presentan los puntos de optimalidad para cada uno de los clasificadores que se hallan en el recubrimiento convexo superior, para las colecciones en español e inglés. En la primera y segunda columnas aparecen los valores de (FPR,TPR) para cada clasificador óptimo. Los clasificadores se identifican por medio de la letra inicial (C – C4.5, P – PART, S – SVM), los meta-esquemas por medio de la letra intermedia (W – pesos, M – MetaCost), y las distribuciones por medio de las finales ([i] + *coste*, siendo i la inversa del coste). Así, "SWc020" representa el clasificador obtenido usando SVM adaptadas al coste con el método de pesos, para una distribución de costes en que los falsos positivos son 20 veces más importantes que los falsos negativos. A la vista de estos datos, es reseñable que:

[4]Exceptuando la combinación de MetaCost con los cuatro algoritmos de aprendizaje, para el español.

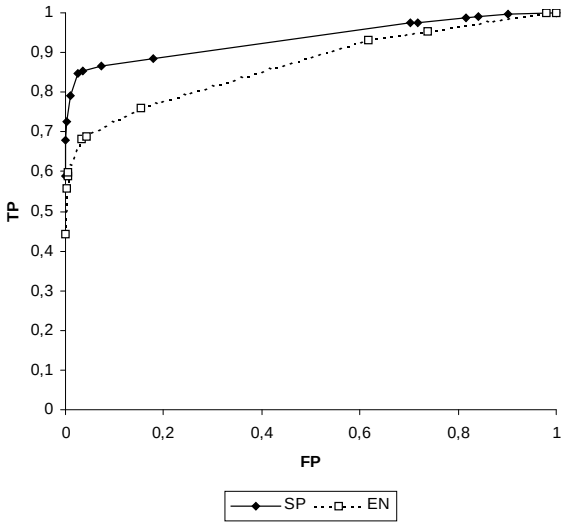

Figura 1: Recubrimiento convexo superior para los enfoques evaluados, en español (SP) e inglés (EN).

Español			**Inglés**		
FPR	**TPR**	**Clasif.**	**FPR**	**TPR**	**Clasif.**
0,000	0,588	SW100	0,000	0,442	SW005
0,001	0,678	SW020	0,002	0,558	SM001
0,002	0,726	SW005	0,005	0,589	SM005
0,010	0,791	SW001	0,006	0,597	SM010
0,025	0,846	SWi005	0,033	0,682	PM010
0,035	0,854	SWi010	0,044	0,690	PM030
0,072	0,866	SWi020	0,153	0,760	PWi060
0,180	0,886	SWi030	0,618	0,930	PWi300
0,702	0,974	PWi080	0,737	0,953	CWi900
0,718	0,976	PWi090	0,981	1,000	PM300
0,815	0,987	SWi300			
0,842	0,990	CWi200			
0,901	0,996	PWi700			
1,000	1,000	CWi400			

Tabla 1: Puntos y clasificadores óptimos para los experimentos en español e inglés.

- Ninguno de los algoritmos y meta-esquemas de sensibilización al coste es claramente dominante para ningún idioma, como suele ocurrir en entornos reales. El algoritmo de bayes ingenuo y el esquema basado en umbrales resultan subóptimos para cualesquiera condiciones operativas.

- Para los experimentos en español, el algoritmo más frecuentemente ganador son las SVM, en combinación con el meta-esquema basado en pesos. Estos resultados son coherentes con los obtenidos para la detección de correo comer-

cial no solicitado en (Gómez, 2002), aunque es preciso resaltar que la combinación con MetaCost no ha sido evaluada en esta colección. Para los experimentos en inglés, es sin embargo PART el algoritmo más frecuentemente óptimo, y MetaCost el meta-esquema más efectivo.

En condiciones extremas o cercanas a ellas, es decir, cuando no se admite ningún falso positivo (página Web pornográfica clasificada como segura), el clasificador más efectivo es SVM con pesos (español) o con MetaCost (inglés). Por ejemplo, con SVM + pesos se alcanza una TPR de 0,588 para el español, lo que significa que se clasifican correctamente un 58,8 % de las páginas seguras cuando no se yerra sobre ninguna pornográfica.

Los porcentajes obtenidos en condiciones extremas son insuficientes para un sistema real, aunque prometedores. Un análisis *post-mortem* de los resultados revela deficiencias en la colección de evaluación. Porcentajes significativos de páginas poseen poco o ningún texto, corresponden a errores de descarga tipo 404, o son páginas basadas en marcos.

6. Conclusiones y trabajo futuro

Aunque los experimentos realizados son prometedores, es preciso avanzar en esta línea, para lo cual nos proponemos: (1) enriquecer y refinar la colección de entrenamiento y evaluación, extrayendo más páginas del ODP e internas de los sitios Web referenciados, eliminando aquellas correspondientes a errores, y extrayendo el contenido de los marcos HTML; y (2) extender los experimentos a otros algoritmos de aprendizaje (en concreto, usando el algoritmo de los k vecinos más cercanos, y el esquema de meta-aprendizaje AdaBoost aplicado a C4.5) y otras representaciones del texto (pesos tipo TF.IDF).

Aunque en la práctica es imposible alcanzar índices de acierto del 100 %, esperamos alcanzar resultados muy cercanos a ellos.

Bibliografía

ACM. 2002. Internet abuse in the workplace. *Communications of the ACM*, 45(1).

Domingos, P. 1999. Metacost: A general method for making classifiers cost-sensitive. En *Proceedings of the 5th International Conference on Knowledge Discovery and Data Mining*.

Frank, E. y I.H. Witten. 1998. Generating accurate rule sets without global optimization. En *Machine Learning: Proceedings of the Fifteenth International Conference*, páginas 144–151. Morgan Kaufmann Publishers.

Gómez, J.M. 2002. Evaluating cost-sensitive unsolicited bulk email categorization. En *Proceedings of the ACM Symposium on Applied Computing*.

Joachims, T. 2001. A statistical learning model of text classification with support vector machines. En *Proceedings of the 24th ACM International Conference on Research and Development in Information Retrieval*. ACM Press.

Lewis, D.D. 1998. Naive (Bayes) at forty: The independence assumption in information retrieval. En *Proceedings of the 10th European Conference on Machine Learning*, páginas 4–15. Springer Verlag.

Provost, F. y T. Fawcett. 2001. Robust classification for imprecise environments. *Machine Learning Journal*, 42(3):203–231.

Quinlan, J.R. 1993. *C4.5: Programs for Machine Learning*. Morgan Kaufmann.

Salton, G. 1989. *Automating text processing: the transformation, analysis and retrieval of information by computer*. Addison-Wesley.

Sebastiani, F. 2002. Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1):1–47.

Ting, K.M. 1998. Inducing cost-sensitive trees via instance weighting. En *Proceedings of the 2nd European Symposium on Principles of Data Mining and Knowledge Discovery*, páginas 139–147.

Witten, I.H. y E. Frank. 1999. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann.

Yang, Y. 1999. An evaluation of statistical approaches to text categorization. *Information Retrieval*, 1(1/2):69–90.

Yang, Y. y J.O. Pedersen. 1997. A comparative study on feature selection in text categorization. En *Proceedings of the 14th International Conference on Machine Learning*.