# INTELLIFILTER - SISTEMA PARA EL FILTRADO DE SITIOS NOCIVOS BASADO EN INTELIGENCIA COMPUTACIONAL

Héctor Cadavid, Diana Rodríguez, Jorge Cely, Juan Pablo Garcia
Escuela Colombiana de Ingeniería

## Introducción

En nuestra sociedad actual, cada vez más influenciada por los nuevos paradigmas de comunicación e interacción social generados por Internet, es ampliamente conocida la problemática con la creación y publicación para libre acceso de contenidos nocivos para menores de edad, como los sitios pornográficos, los sitios que incitan a la violencia, e incluso sitios que ponen en riesgo la integridad física de los niños al ponerlos en contacto con explotadores de menores.

A causa del principio de libertad de expresión, ha sido imposible que alguien tome el control de estos contenidos, por lo cual, la responsabilidad de proteger a los menores de esta exposición recae ahora en los padres y en las instituciones educativas. Es por esto, que desde hace varios años, la pregunta de cómo crear una herramienta capaz de detectar con suficiente precisión si un sitio es nocivo o no, se ha vuelto una cuestión de investigación abierta de mucha importancia.

## Descripción del problema

Como un esfuerzo para proveer a padres y educadores de herramientas que permitan convertir las aulas de informática y los computadores del hogar en recursos seguros para los menores, desde la industria y desde la academia se han creado filtros de contenidos, que en su mayoría se basan en el esquema de coincidencia de palabras clave (se determina si un sitio es nocivo si contiene determinadas palabras), el cual ha tenido siempre el problema de los 'falsos positivos' con páginas de contenidos no nocivos (el ejemplo tradicional, es el constante bloqueo de sitios relacionados con anatomía humana, que se confunden con sitios pornográficos).
También se han creado sistemas basados en 'listas negras', donde simplemente se tiene una base de datos de los sitios reportados como nocivos. Esta solución sin embargo, ha resultado poco efectiva, dado el hecho de que es inviable mantener actualizadas estas listas con los miles de sitios Web emergentes día a día.

Por otro lado, las técnicas de inteligencia computacional resultan ser muy apropiadas para los problemas de clasificación de información desconocida a partir de ejemplos conocidos -lo cual constituye el problema en mención-, ya que permiten, a partir de un número grande de muestras, determinar características no evidentes para una clasificación más precisa. Por esto, se propone hacer uso de dichas técnicas como aproximación a la solución del problema de clasificación y filtrado de contenidos Web.

## Alcance

El alcance de este trabajo consta de dos elementos principales: un Proxy de Internet, y un conjunto de clasificadores que constituyen el criterio de filtrado.
El Proxy de Internet, se encarga de interceptar las comunicaciones entre el navegador de un computador y los servidores de Internet, para determinar si un sitio solicitado por el usuario es nocivo o no, mediante el uso de los clasificadores integrados a él. Estos clasificadores están basados en técnicas de inteligencia computacional supervisados, es decir, que requieren de un entrenamiento previo, para nuestro caso, con muestras de páginas clasificadas previamente como nocivas o no nocivas.

## Metodología

*Metodología de desarrollo*

Se hizo uso de la metodología ágil XP (Programación Extrema), haciendo énfasis en los ciclos de desarrollo cortos, y en el desarrollo orientado por pruebas.

## Fuentes de información

La principal fuente de información fueron libros y artículos de minería de datos y aprendizaje de máquina, junto con los tutoriales disponibles libremente en Internet (ver principales referencias en la sección de bibliografía).

## Descripción del producto

El producto desarrollado puede describirse en las dos etapas del proceso de puesta en funcionamiento del filtro: el entrenamiento y la clasificación. En esta primera versión, se incorporan los siguientes modelos de clasificación:

- Redes neuronales
- *Naive* Bayes
- Árboles de decisión

*Proceso de entrenamiento*

Para el proceso de entrenamiento, se parte de dos conjuntos de URL. Uno, de sitios previamente catalogados como nocivos (particularmente páginas con contenido pornográfico), y otro de sitios que tengan reconocido ser no nocivos. A partir de esta información, se realiza el siguiente proceso (descrito en la figura 1):

1. Se extrae el texto contenido en cada URL, removiendo palabras de parada (pronombres, conectores y otras palabras auxiliares) de acuerdo al idioma identificado para dicho contenido, en caso de que dicho URL siga en línea. Inicialmente el producto reconoce dos idiomas, inglés y español.
2. Para el contenido extraído del conjunto completo en cada categoría y para cada idioma, se registra en una base de datos la información estadística de las palabras contenidas en él, específicamente el número de ocurrencias de cada palabra.
3. A partir de la información estadística obtenida de las páginas, se aplica una función que determina qué palabras harán parte del vector de características de las páginas. Un vector de características será un arreglo de bits, donde cada posición corresponde a una palabra, y su valor, para un URL determinado, será 1 ó 0 dependiendo de si existe o no dicha palabra en su contenido, teniendo en cuenta que una palabra se considera existente dentro de un texto, si cumple con una frecuencia mínima establecida.
4. Una vez se determine cuál será el vector de características, se calculan los vectores de cada una de las contenidos extraídos inicialmente.
5. Los contenidos de cada uno de los URL de los conjuntos iniciales, transformados a su forma de vector binario, son aplicados a un proceso de entrenamiento para cada una de las técnicas que vaya a utilizar el filtro integrado al proxy.
6. Los datos resultantes de estos procesos de entrenamiento se hacen persistentes, de modo que con éstos las distintas técnicas de clasificación puedan aplicarse a nuevos contenidos, sin tener que repetir el entrenamiento.

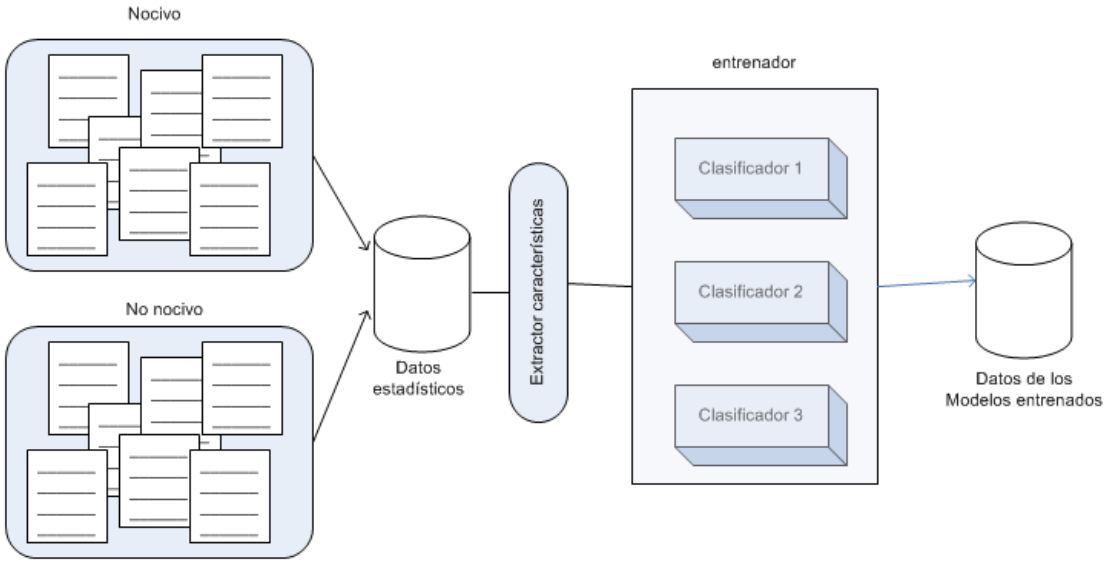


Figura 1: Proceso de entrenamiento de los clasificadores.

*Proceso de interceptación y clasificación*

La herramienta de filtrado, como se mencionó anteriormente, se basa en el uso de un Proxy de Internet, que intercepta el tráfico del navegador hacia y desde Internet. Éste, como se muestra en la figura 2, una vez captura el contenido de un sitio de Internet, extrae su contenido textual, y lo transforma al vector de características descrito anteriormente. Éste vector de características es dado al clasificador previamente entrenado, y dependiendo de la clasificación dada por este último (no nocivo o nocivo), se le envía al usuario, bien sea el contenido original, o un contenido alterno sano.

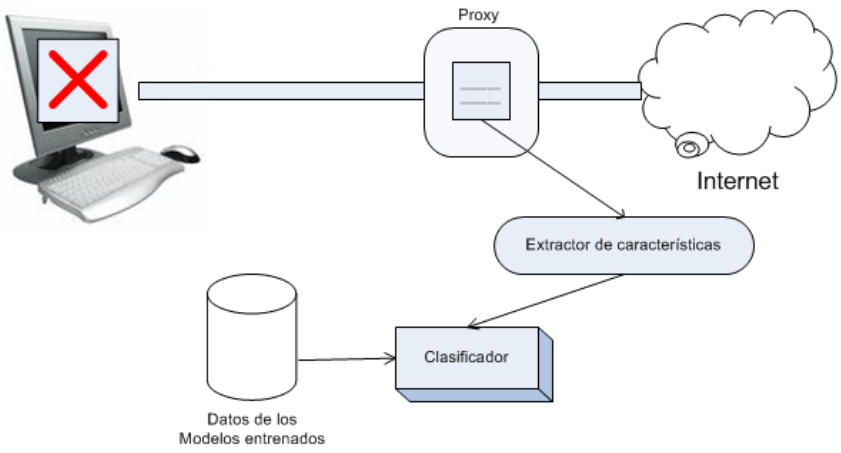


Figura 2: funcionamiento del Proxy basado en los clasificadores

**Experimentación y resultados obtenidos**

Para el proceso de experimentación, se hizo uso de una lista de sitios de Internet, previamente categorizados, que contiene cerca de dos millones de direcciones. A partir de esta lista, se configuraron experimentos para las técnicas *naive Bayes*, redes neuronales, y árboles de

decisión, variando en número de muestras y el tamaño de los vectores de características. En todos los casos, el 90% de los datos fue utilizado para entrenamiento, y el 10% para validación.

*Resultados obtenidos*

Una vez realizado el proceso de entrenamiento para las tres técnicas, los siguientes fueron los resultados del proceso de validación:

1. *NAYBE VAYES*

| *# características* | *Tamaño de la muestra (# de páginas)* | *Matriz de confusión* | *Precision* | *Exhaustividad* |
|---|---|---|---|---|
| *100* | *1000* | a b <-- classified as<br>492 8 \| a = 0<br>142 358 \| b = 1 | *0,97* | *0,71* |
| *100* | *5000* | a b <-- classified as<br>2435 65 \| a = 0<br>746 1754 \| b = 1 | *0,96* | *0,70* |
| *100* | *9000* | a b <-- classified as<br>4374 126 \| a = 0<br>1481 3019 \| b = 1 | *0,95* | *0,67* |
| *1000* | *1000* | a b <-- classified as<br>498 2 \| a = 0<br>139 361 \| b = 1 | *0,99* | *0,72* |
| *1000* | *5000* | a b <-- classified as<br>2480 20 \| a = 0<br>700 1800 \| b = 1 | *0,98* | *0,72* |

2. *ÁRBOLES DE DECISIÓN*

| *# características* | *Tamaño de la muestra (# de páginas)* | *Matriz de confusión* | *Precision* | *Exhaustividad* |
|---|---|---|---|---|
| *100* | *1000* | a b <-- classified as<br>488 12 \| a = 0<br>115 385 \| b = 1 | *0,96* | *0,77* |
| *100* | *5000* | a b <-- classified as<br>2403 97 \| a = 0<br>529 1971 \| b = 1 | *0,95* | *0,78* |
| *100* | *9000* | a b <-- classified as<br>4349 151 \| a = 0<br>1090 3410 \| b = 1 | *0,95* | *0,75* |
| *1000* | *1000* | a b <-- classified as<br>495 5 \| a = 0<br>58 442 \| b = 1 | *0,98* | *0,88* |
| *1000* | *5000* | a b <-- classified as<br>2428 72 \| a = 0<br>227 2273 \| b = 1 | *0,96* | *0,90* |

*3. REDES NEURONALES*

| *# características* | *Tamaño de la muestra (# de páginas)* | *Matriz de confusión* | *Precision* | *Exhaustividad* |
|---|---|---|---|---|
| *100* | *1000* | `a b <-- classified as`<br>`470 30 \| a = 0`<br>`64 436 \| b = 1` | *0,93* | *0,87* |
| *100* | *5000* | `a b <-- classified as`<br>`2426 74 \| a = 0`<br>`498 2002 \| b = 1` | *0,96* | *0,80* |
| *100* | *9000* | `a b <-- classified as`<br>`4432 68 \| a = 0`<br>`1375 3125 \| b = 1` | *0,97* | *0,69* |

*Análisis de los resultados*

Como se observa en las tablas anteriores, todas las técnicas, en general, tienen un alto grado de precisión. Esto significa, que de todas las páginas identificadas como nocivas, un alto porcentaje son, en efecto, nocivas.
Por otro lado, la exhaustividad, se refiere a la proporción de sitios identificados como nocivos respecto a la totalidad de la muestra. Esta métrica, aunque fue alta, no lo fue tanto como la precisión, y requerirá del uso de técnicas complementarias de aprendizaje de máquina como aquellas basadas en *boosting*, que permiten la integración de varios clasificadores para construir uno solo de mejor desempeño.

En general los resultados son prometedores, sobretodo teniendo en cuenta que aún no se ha integrado el procesamiento de imágenes, y no se han incorporado heurísticas más desarrolladas de selección óptima de características (como por ejemplo, la selección por entropía).

**Aplicaciones**

- Herramienta de filtrado para el hogar con un solo computador. Requeriría de la instalación de un software y la configuración del navegador en el computador.
- Herramienta de filtrado institucional o en redes caseras. Esta solución, más robusta, requeriría la configuración de un servidor de salida a Internet.
- Herramienta complementaria a servidores DNS. Un servidor DNS podría, a medida que entregue direcciones, verificar su contenido con la herramienta, e ir restringiendo las direcciones que resulten nocivas.

**Limitaciones**

- En esta etapa inicial, el filtro requiere realizar un entrenamiento diferente para cada idioma, con lo que no se estarían controlando los contenidos de sitios en idiomas que no hayan sido contemplados
- Existe una degradación de desempeño en la respuesta del navegador que depende del clasificador. Esto requiere del afinamiento de los parámetros del clasificador, o incluso de su implementación para mejorar los tiempos de respuesta.

**Ventajas**

- La herramienta podrá determinar la categoría nociva o no nociva de un sitio en Internet, así este sea nuevo y nadie lo haya categorizado previamente.
- Es transparente para el usuario, pues la funcionalidad de su navegador será exactamente la misma, salvo si intenta acceder a contenidos nocivos.

- Puede extenderse para manejar más de dos categorías, con lo que el perfil de navegación podrá ajustarse a rangos de edades más específicos.

**Trabajo futuro**

- Integrar un esquema de integración de técnicas, para mejorar la precisión de la clasificación a través del uso conjunto de varios clasificadores.
- Extender el modelo para poder clasificar más de dos categorías.
- Agregar técnicas de procesamiento de imágenes para filtrar también las fotos disponibles en las páginas.

**Bibliografía**


[1]. R. Agrawal. Tutorial: Data mining. In ACM, editor, 13th Symposium - 1994 May: Minneapolis; MN, volume 13 of PROCEEDINGS OF THE ACM SIGACT SIGMOD SIGART SYMPOSIUM ON PRINCIPLES OF DATABASE SYSTEMS 1994, pages 75-76, New York, NY 10036, USA, 1994. ACM Press.

[2]. Rakesh Agrawal. Data mining. In Proceedings of the 13th Symposium on Principles of Database Systems, pages 75-76, New York, NY, USA, May 1994. ACM Press.

[3]. David Aha. Machine learning tutorial (slides and anotated bibliography).

[4]. Fabrizio Sebastiani. Machine Learning in Automated Text Categorization.

[5]. P. Cheeseman and J. Stutz. Bayesian classification (AUTOCLASS): Theory and results. In U. M. Fayyad, G. Piatetsky-Shapiro, P Smyth, and R. Uthurusamy, editors, Advances in Knowledge Discovery and Data Mining. 1995.

[6]. H. J. Lu, R. Setiono, and H. Liu. Effective data mining using neural networks. Ieee Trans. On Knowledge And Data Engineering, 8:957-961, 1996.

[7]. Michel Manago and Yves Kodrato. Induction of decision trees from complex structured data. In Gregory Piatetsky-Shapiro and William J. Frawley, editors, Knowledge Discovery in Databases, pages 289-306. AAAI Press / The MIT Press, Menlo Park, California, 1st edition, 1991.

[8]. Berndt Müller and Joachim Reinhardt. Neural Networks, an introduction. Physics of Neural Networks. Springer-Verlag, Berlin, 1991.