Revista de la Red Colombiana de Programas de Ingeniería de Sistemas y Afines, Volumen 4, año 4, No. 4. ISSN 2256-2567, Deciembre de 2011

# Intellifilter: sistema de filtrado parental soportado por aprendizaje maquinal supervisado

Héctor Fabio Cadavid Rengifo
Jorge Humberto Cely Higuera
Juan Pablo García Segura

hector.cadavid@escuelaing.edu.co, {jorge.cely, juan.garcia}@mail.escuelaing.edu.co
Escuela Colombiana de Ingeniería

*Resumen*- En este artículo se presenta la herramienta *Intellifilter* para el filtrado de contenidos de internet no aptos para niños, basada en un conjunto de técnicas de aprendizaje supervisado[1] para clasificación de texto e imágenes, junto con una infraestructura de interceptación de peticiones HTTP basada en *software*. Cuando *Intellifilter* esté habilitado en un equipo o red en particular, todos los contenidos solicitados mediante el protocolo HTTP serán evaluados por un conjunto de clasificadores previamente entrenados y, de acuerdo con el resultado de dicha categorización, se aceptará o no su entrega al destinatario original. Aunque la herramienta aún se encuentra en una fase 'beta', los resultados experimentales obtenidos con las técnicas utilizadas muestran resultados promisorios que indican que éste puede ser un buen punto de partida para el desarrollo de mecanismos de control de los contenidos de internet.

Palabras claves: filtro web, inteligencia computacional, árboles de decisión, redes neuronales, *naive bayes*, clasificación, *proxy*, *browser*.

[1] Las técnicas de aprendizaje maquinal supervisado son métodos para categorizar información a partir de un conjunto de ejemplos previamente clasificados. Con éstos, y mediante un proceso de entrenamiento, dichas técnicas se aproximarán a las reglas que permitirán clasificar automáticamente nuevos conjuntos de datos [13].

## I. INTRODUCCIÓN

La pornografía, presente en la sociedad desde hace décadas (si no siglos), con el auge de internet se ha convertido en un fenómeno intrusivo y omnipresente, ya que los computadores en red, presentes prácticamente en cualquier institución u hogar, son una ventana abierta a un volumen extremadamente grande de este tipo de contenidos (aproximadamente el 12% del total de sitios de internet[2]). Esto resulta especialmente grave para los niños, quienes desde temprana edad quedan expuestos a estos contenidos que -según pruebas científicas- los puede afectar negativamente (3,7).

Por la naturaleza descentralizada y libre de internet, no se ve factible en el futuro cercano una solución radical a este gran problema. Salvo por la normatividad contra la pornografía infantil, existente en casi todos los países del mundo, los contenidos pornográficos 'convencionales' siempre serán legales y accesibles para todo el público, incluyendo el infantil. Por esta razón, y en vista de que la prevención de los problemas antes mencionados queda prácticamente en manos de los padres y las instituciones educativas, es urgente crear un mecanismo confiable que permita garantizar que los contenidos a los que acceden los niños a su cargo sean seguros.

En este artículo se describe la herramienta *Intellifilter*, la cual busca aproximarse como solución para proteger contra contenidos no aptos a los menores en los puntos de acceso a internet utilizados frecuentemente por ellos: el hogar y las instituciones académicas.

[2] De acuerdo con el sitio *Healty Mind*: http://www.healthymind.com/s-porn-stats.html.

## II. ESTADO DEL ARTE

El conjunto de aproximaciones al problema de filtrar los contenidos a los que se expone un usuario de internet se encuentra en dos medios diferentes: el académico y el comercial. En este último se puede encontrar un grupo de productos –no muy grande- capaz de controlar, además de los contenidos, los sistemas de internet potencialmente peligrosos para los niños, como los chats, las redes sociales y otros servicios como *Usenet* y FTP. Este tipo de herramientas presentan funcionalidades sofisticadas como control de tiempo de conexión a internet y notificación de conductas irregulares vía correo electrónico o mensajes de texto, lo que permite una supervisión que involucra a los padres.

Entre las herramientas más populares de los 'escalafones' de *software* hechos en sitios especializados en internet[3], sobresalen *NetNanny* y Cyberpatrol, tanto por su relativo buen desempeño como por las opciones de vigilancia y control que ofrecen a los padres. De éstos se puede resaltar:

**NetNanny:** filtra los sitios al que acceden los usuarios de acuerdo con un análisis inmediato sobre su contenido, que se basa en el estudio estadístico de la ocurrencia de ciertas palabras distintivas de los contenidos no aptos para menores, junto con una técnica denominada '*Dynamic contextual analysis*'. Con la técnica de "análisis dinámico en contexto", se verifica si dichas "palabras claves" encontradas en un sitio web son usadas o no en un ambiente realmente no apto para menores, para evitar el bloqueo de sitios aptos que usan algunas de estas palabras de forma adecuada (por ejemplo los dedicados a temas de salud).

[3] Sitios como http://www.consumersearch.com/parental-control-software.

Esta herramienta requiere la instalación de un *software* en cada equipo que se quiera proteger, y permite la administración remota de la configuración del mismo (14).

**Cyberpatrol:** filtra los contenidos de internet de acuerdo con una 'lista negra' (privada) de sitios web, que es actualizada de forma continua mediante una tecnología denominada 'SiteCat', la cual funciona de forma similar a los motores de búsqueda de internet, pues recorre de forma exhaustiva los sitios disponibles en la red (mediante un *crawler*), analiza su contenido y los clasifica en la lista antes mencionada. Al igual que NetNanny, debe ser instalada en cada equipo que requiera control de contenidos (15).

En el medio comercial existen también soluciones que no necesitan instalación de *software*, ya que funcionan a más bajo nivel sobre la infraestructura de internet. Tal es el caso de *OpenDNS – Family Shield*, un servidor de nombres de dominio (DNS/*Domain Name Service*) ajustado de manera que sólo resuelve aquellos que no se encuentren en una lista negra previamente construida.

Por otro lado, en el medio académico (es decir, en la literatura científica disponible), se pueden encontrar principalmente propuestas basadas en técnicas de aprendizaje maquinal supervisado para resolver partes del problema del filtrado de contenidos tales como la clasificación de imágenes (8, 10, 11) y de materiales textuales (4, 5). Adicionalmente, en la literatura se pueden encontrar trabajos enfocados al problema de la interceptación eficiente de los contenidos para su posterior procesamiento (2).

Las técnicas en mención, aplicadas al problema de clasificación de textos e imágenes, resultan muy apropiadas para efectos de filtrar contenido, ya que permiten, a partir de sitios web previamente definidos, clasificar los desconocidos. Con esto, en principio, ya no son necesarias las listas negras ni las dispendiosas y costosas tareas de clasificación manual de contenidos. Además, la tasa de falsos positivos puede ser controlada ajustando los parámetros de los algoritmos utilizados.

### Intellifilter respecto de las herramientas de filtrado existentes

Aunque las herramientas de filtrado de contenidos existentes en el mercado cumplen bien con su propósito de evitar que los menores accedan a contenidos poco adecuados en internet, Intellifilter se propone como una aplicación que, teniendo el mismo fin, presenta los siguientes valores agregados:

- Al igual que soluciones como OpenDNS, no necesita instalación de *software* en los PC que requieren filtrado de contenidos, ya que se instala a manera de servidor *proxy* de internet. Sin embargo, a diferencia de OpenDNS, Intellifilter realiza un filtrado de contenidos de forma dinámica, por lo que su efectividad no depende de que el sitio visitado exista o no en una lista negra.
- Al no requerir instalación de *software* en el PC que se desea proteger, es menos propenso a otras “trampas” de los usuarios para saltar el proceso de filtrado (por ejemplo, manipular el sistema operativo para interrumpir el proceso responsable del filtrado).

- Al ser un *software* abierto, con una arquitectura por componentes y orientada a la integración de cualquier técnica de clasificación de textos e imágenes, quien así lo quiera podrá mejorar el desempeño de Intellifilter mediante la implementación de dicho clasificador.

## III. Arquitectura de Intellifilter

Teniendo en cuenta el gran potencial de las técnicas de aprendizaje supervisado para la clasificación de contenidos desconocidos, *Intellifilter* se propone como una herramienta de filtrado de contenidos basada exclusivamente en el criterio de clasificadores previamente entrenados, a la cual se le puede integrar cualquier tipo de clasificador (siempre y cuando sea supervisado).

## IV. Clasificación de contenidos textuales

Como se mencionó anteriormente, *Intellifilter* está soportado por un conjunto de clasificadores previamente entrenados para determinar si un contenido es apto para menores o no. En cuanto a la clasificación de contenidos textuales, la figura 1 presenta la estrategia de entrenamiento desarrollada para los clasificadores:

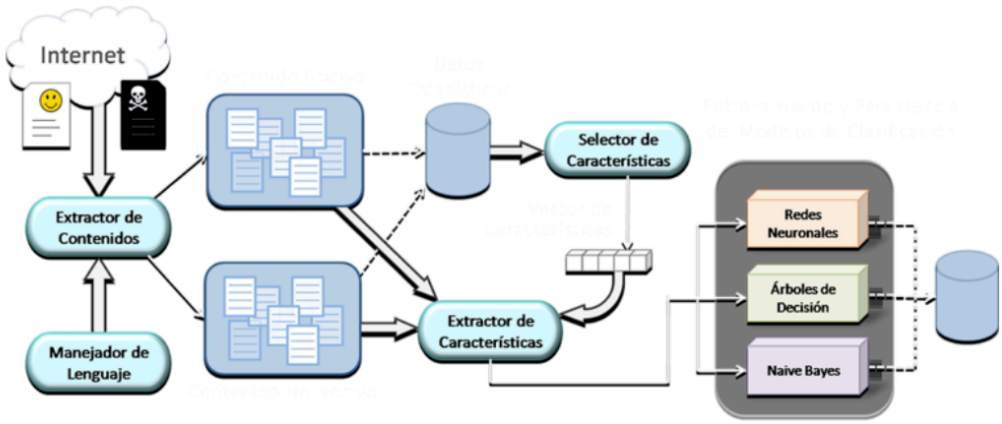


Figura 1. Arquitectura del módulo de entrenamiento de Intellifilter.

Esta estrategia, que parte de dos conjuntos de URL, uno de sitios previamente catalogados como no aptos para menores (particularmente páginas con contenido pornográfico), y otro de sitios reconocidos como aptos para todo el mundo, realiza el siguiente proceso:

1. Se extrae el texto contenido en cada URL, removiendo palabras de parada (pronombres, conectores y otras auxiliares) de acuerdo con el idioma identificado.

2. Para el contenido extraído del conjunto completo en cada categoría (aptos/no aptos para menores) y para cada idioma, se registra en una base de datos la información estadística de las palabras halladas en él.

3. A partir de la información estadística obtenida de las páginas, se aplica una función que determina cuáles harán parte del vector de características[4] de las páginas.

4. Una vez determinado el vector de características, se calculan los vectores de cada uno de los contenidos extraídos inicialmente. Dichos vectores son aplicados a un proceso de entrenamiento para cada técnica que vaya a utilizar el filtro.

6. Para poder obtener indicadores de precisión y exhaustividad, se entrenan diferentes modelos con cada técnica escogidas para la clasificación.

[4] Es un vector de bits en el que cada posición corresponde a la ocurrencia o ausencia de una palabra dentro de un contenido textual.

## V. Proceso de interceptación, clasificación y filtrado

Una vez se han realizado las tareas de entrenamiento de los clasificadores de *Intellifilter* (con aquellos parámetros que hayan arrojado un mejor desempeño), la herramienta queda lista para instalarse, a manera de servidor *proxy*, bien sea en un computador personal o en el servidor de una intranet.

Como se muestra en la figura 2, una vez el *proxy* de *Intellifilter* intercepta el contenido de un sitio de internet, extrae su contenido textual y lo transforma al vector de características descrito anteriormente. Este vector de características es dado al clasificador previamente entrenado, y dependiendo de la clasificación dada por este último (apta o no apta para menores), se le envía al usuario, bien sea el contenido original, o uno predefinido que informa sobre la restricción del sitio visitado.

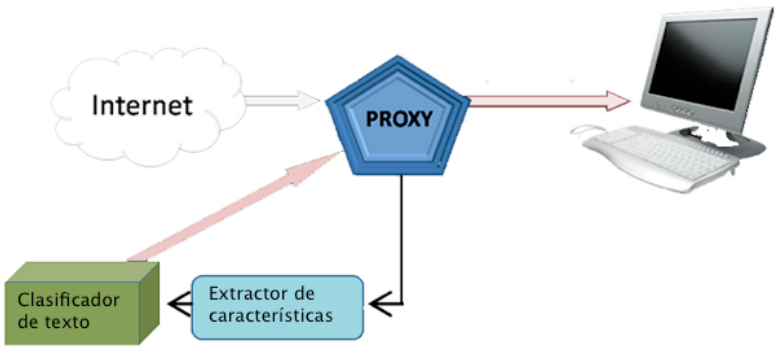


Figura 2. Proceso de interceptación mediante un proxy.

## VI. Experimentación y resultados

El proyecto Intellifilter contempló la implementación de tres diferentes clasificadores de texto, con el fin de identificar cual tenía un mejor desempeño, y así usarlo por defecto en el producto: naive Bayes, árboles de decisión y redes neuronales. Dado que cada una de estas técnicas tiene un desempeño diferente

de acuerdo con el número de características usadas y de datos de entrenamiento, para identificar cuál se comportaba mejor se montó un proceso de ejecución automática de experimentos.

### Resultados obtenidos

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

Naive Bayes

| # características | Tamaño de la muestra (# de páginas) | Precisión | Exhaustividad |
|---|---|---|---|
| 100 | 1000 | 0,97 | 0,71 |
| 100 | 5000 | 0,96 | 0,70 |
| 100 | 9000 | 0,95 | 0,67 |
| 1000 | 1000 | 0,99 | 0,72 |
| 1000 | 5000 | 0,98 | 0,72 |

Árboles de decisión

| # características | Tamaño de la muestra (# de páginas) | Precisión | Exhaustividad |
|---|---|---|---|
| 100 | 1000 | 0,97 | 0,71 |
| 100 | 5000 | 0,96 | 0,70 |
| 100 | 9000 | 0,95 | 0,67 |
| 1000 | 1000 | 0,99 | 0,72 |
| 1000 | 5000 | 0,98 | 0,72 |

Redes neuronales

| # características | Tamaño de la muestra (# de páginas) | Precisión | Exhaustividad |
|---|---|---|---|
| 100 | 1000 | 0,93 | 0,87 |
| 100 | 5000 | 0,96 | 0,80 |
| 100 | 9000 | 0,97 | 0,69 |

## VII. 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 clasificadas por la herramienta como no aptas para menores, un alto porcentaje lo son en la realidad.

Por otro lado, la exhaustividad, aunque se aprecia relativamente alta, no lo fue tanto como la precisión. Este resultado sugiere la necesidad de experimentar con alternativas para el mejoramiento del desempeño de los clasificaciores como el *boosting*, que permite la integración de varios clasificadores para construir uno solo de mejor desempeño.

De acuerdo con estos resultados, el clasificador idóneo, según la arquitectura presentada, fue el árbol de decisión usando 1.000 características. Sin embargo, es posible que los demás clasificadores sigan siendo útiles, ya que el 4% y 10% de error en cuanto a precisión y exhaustividad, respectivamente, podrían eventualmente mejorarse ponderando las clasificaciones de los tres clasificadores mediante las técnicas de *boosting* indicadas anteriormente.

## TRABAJO FUTURO

Como trabajo futuro, se plantean los siguientes objetivos:

- Incorporar y experimentar con algoritmos de selección de características para la clasificación de textos. En este caso, estos algoritmos permitirán identificar el conjunto de palabras más representativo para todo el grupo de datos de entrenamiento y, por lo tanto, debería ayudar a obtener mejores resultados en los experimentos.

- En esta primera versión, Intellifilter controla el acceso a los contenidos de internet mediante el análisis de los contenidos textuales. Con el fin de aumentar la precisión, en un futuro se espera incorporar también la clasificación, en tiempo real, de las imágenes recibidas por el cliente desde internet.

## CONCLUSIONES

- El enorme volumen de pornografía disponible abiertamente en internet para cualquier tipo de público (incluyendo niños), es un problema que hasta

ahora está manifestando sus efectos, en enfermedades como la adicción o en comportamientos negativos como la violencia sexual. Sin embargo, tarde o temprano la sociedad tendrá que caer en la cuenta de que debe tomar cartas en el asunto si no quiere que este problema alcance mayores proporciones. Por esta razón, el aporte que se puede hacer desde la tecnología, aunque no sea una solución definitiva, contribuirá al menos a que los hogares y centros educativos no sean foco de la diseminación de este problema.

- Teniendo en cuenta las características de Intellifilter indicadas anteriormente, como su facilidad de mejoramiento y de incorporación a una infraestructura de red, junto al hecho de que los resultados obtenidos se dieron con algoritmos convencionales (en el medio académico están en desarrollo otros más sofisticados y de mejor desempeño), la potencialidad de esta herramienta, viéndola como un medio de aplicación concreto de los trabajos que se desarrollan en el área de aprendizaje de máquina, es muy alta.

## REFERENCIAS


[1]. Adams, R. and Bischof, L. (1994). Seeded region growing. Pattern analysis and machine intelligence. *IEEE Transactions on*, 16(6), pp. 641-647.

[2]. Akbas, E. (2008, October). Next generation filtering: oine filtering enhanced proxy architecture for web content filtering. pp. 1-4.

[3]. Cline, V.B. (1990). Pornography's effects on adults and children. New York: Morality in Media, 11.

[4]. Deselaers, T., Pimenidis, L. and Ney, H. (2008, December). Bag-of-visual-words models for adult image classification and filtering. pp. 1 -4.

[5]. Youngsoo, K. and Taekyong, N. (2006, February). An efficient text filter for adult web documents. 1, p. 3.

[6]. Kotsiantis, S.B. (2007). Supervised machine learning: a review of classification techniques. *Informática*, 31(3), pp. 249-268.
[7]. Eberstadt, M. and Layden, M.A. (2010). *The social costs of pornography: a statement of findings and recommendations*. (1st. ed.). The Witherspoon Institute.
[8]. Xuanjing, Shen, Wei Wei, Quingji Qian. (2997). The filtering of internet images based on detecting erotogenic part. In ICNC '07: Proceedings of the Third International Conference on Natural Computation, pages 732-736. Washington, D.C., USA. IEEE Computer Society.
[9]. Varadharajan, V. (2010, July). Internet filtering issues and challenges. *Security Privacy*, IEEE, 8(4), pp. 62-65.
[10]. Yi-Ding, W. and Jing-Nan, G. (2009, July). A method of erotic images filtering in real internet, 3, pp. 1477-1481.
[11]. Qing-Fang, Z. and Wei Zeng, G.W. (2004). Shape-basedadult image detection. In ICIG '04: Proceedings of the Third International Conference on Image and Graphics, pp. 150-153. Washington, D.C., USA. IEEE Computer Society.
[12]. The free dictionary. URL: http://encyclopedia2.thefreedictionary.com/Internet+proxy.
[13]. Alpaydin, E. (2004). *Introduction to machine learning.* MIT Press.
[14]. Sitio web de Ensing Solutions, consultor en IT y proveedor de NetNanny, URL: http://contentprotect.co.uk/dca.html.
[15]. Sitio web de CyberPatrol. URL: http://www.cyberpatrol.com/research/sitecat.asp.

**Héctor Fabio Cadavid Rengifo.** Ingeniero de Sistemas de la Escuela Colombiana de Ingeniería y Ms.C. en Ingeniería de Sistemas y computación de la Universidad Nacional de Colombia. Profesor Asistente de la facultad de Ingeniería de Sistemas de la Escuela Colombiana de Ingeniería.

**Jorge Humberto Cely Higuera.** Estudiante de último semestre del programa de Ingeniería de la Escuela Colombiana de Ingeniería.

**Juan Pablo García Segura.** Estudiante de último semestre del programa de Ingeniería de la Escuela Colombiana de Ingeniería.