Software para el filtrado de páginas web pornográficas basado en el clasificador KNN - UDWEBPORN

Software for filtering pornographic web pages using KNN classifier – UDWEBPORN

Jorge E. Rodríguez R, MSc., Harry A. Barrera F. Ing. & Sandra P. Bautista M. Ing.
Grupo de Investigación en Inteligencia Artificial – Universidad Distrital "Francisco José de Caldas"
jrodri@udistrital.edu.co, harrybarrer@gmail.com, sandra_bautista@hotmail.com

Recibido para revisión 02 de septiembre de 2010, aceptado 03 de enero de 2011, versión final 05 de febrero de 2011

***Resumen*— En este artículo se propone el uso de un algoritmo para filtrar automáticamente páginas web, en este caso se filtra páginas pornográficas. Para llevar a cabo dicha (filtrado automático) tarea, se implementan técnicas de minería de datos y algoritmos de aprendizaje incremental para el proceso de extracción, representación y clasificación de las páginas.**

***Palabras Clave*— Aprendizaje Computacional, Clasificación de Hipertexto, Filtrado de Contenido web, Minería web.**

***Abstract*— - In this paper we show the used algorithm for filtering web pages. The paper is focused at development of software to automatically filter web pages, pornographic web pages in this case. To carry out this (automatic filtering), are implemented data mining techniques and incremental learning algorithms for the extraction, representation and classification process.**

***Keywords*— Hypertext Classification, Machine Learning, Web Content Filtering, Web Mining.**

I. INTRODUCCIÓN

Las tecnologías actuales orientadas al filtrado o la restricción de diferentes tipos de contenido web requieren de una constante actualización por parte de personal capacitado (Administradores de Red, Administradores de Sistemas) para su funcionamiento. Las herramientas de software destinados a tal propósito se han encargado hasta el momento de restringir o filtrar el acceso a determinadas páginas definidas previamente de acuerdo a los intereses en cada caso, normalmente mediante listas de control de acceso.

Actualmente, existe en el mundo cerca de 238'000.000 de sitios web [18], que albergan, según informe de uno de los buscadores más reconocidos como es Google, cerca de un trillón (1.000.000.000.000) de URL's únicas. Colombia se encuentra en el tercer lugar de países latinoamericanos con mayor registro de usuarios en Internet frente al total de su población; el último reporte actualizado de la Internet World Stats, indica que el 47,6% de la población accede a la web [11]. Del total de sitios web existentes, el 12% corresponde a páginas pornográficas, sin contar que el 25% de las búsquedas que se realizan y el 35% de las descargas de la web son del mismo tipo, además se estima que cada día aparecen casi 300 nuevos sitios web [21].

De acuerdo a esto se ve la necesidad de desarrollar mecanismos automáticos que no dependan de una constante administración y actualización. En este artículo se plantea el desarrollo de un software para clasificación de hipertexto específicamente pornográfico, utilizando la metodología de Minería de datos CRIPS-DM (http://www.crisp-dm.org/) que puede adaptarse a proyectos de Minería Web. Se implementa una técnica de aprendizaje computacional para realizar el filtrado de las páginas web automáticamente y que funciona como un proxy HTTP (HyperText Transfer Protocol) que intercepta todas las peticiones del navegador y las clasifica. Para este desarrollo se decidió trabajar con páginas en inglés; según estadísticas de la firma Internet World Stats, este es el idioma más popular en Internet con un 39.5% frente al resto de idiomas; y según el número de usuarios que consultan información en un idioma especifico, el inglés registra 499 millones, ocupando el primer lugar frente al mandarín o el español que también son bastante utilizados [12]. Los procedimientos utilizados puedan ser implementados en cualquier otro idioma y dominio aplicando la misma metodología.

Revista Avances en Sistemas e Informática, Vol.8 No.1, marzo de 2011 - Medellín. ISSN 1657-7663

El artículo se encuentra estructurado siguiendo una metodología y procedimiento tanto de desarrollo de software como del filtrado de páginas, adaptada para el desarrollo propuesto, así: - en la primera sección se describe la arquitectura del sistema y las técnicas de extracción y representación de las páginas de acuerdo con el tipo e idioma seleccionados, - en seguida se define la técnica de clasificación utilizada, - posteriormente, se muestran las pruebas realizadas y los resultados obtenidos, - y por último se detallan las conclusiones y propuestas para trabajos futuros que puedan mejorar el desempeño del software o incluir nuevas características en este.

II. ARQUITECTURA DEL SOFTWARE DE FILTRADO

El filtrado de contenido web corresponde al proceso que permite restringir o permitir el acceso a un documento HTML (HyperText Markup Language), basado en algún tipo de análisis que se realiza sobre este. La arquitectura del sistema de filtrado está basada en la propuesta realizada en [5] y se puede observar en la Figura 1. En general la arquitectura del software cuenta con cuatro componentes principales:

Proxy HTTP: se encarga de interceptar las peticiones del navegador, disparar el proceso de clasificación y permitir o denegar la visualización de la página solicitada. Se utiliza la herramienta Paw 0.3 (http://pawproject.sourceforge.net/).

Parser HTML y diccionario: el parser (analizador sintáctico) HTML en conjunto con el diccionario se encargan de extraer la información de una página web. La herramienta utilizada es Jericho 2.6 (http://sourceforge.net/projects/jerichohtml/).

Representación del documento: este componente permite representar una página web como una instancia WEKA (http://www.cs.waikato.ac.nz/ml/weka/) utilizando el API que esta herramienta proporciona.

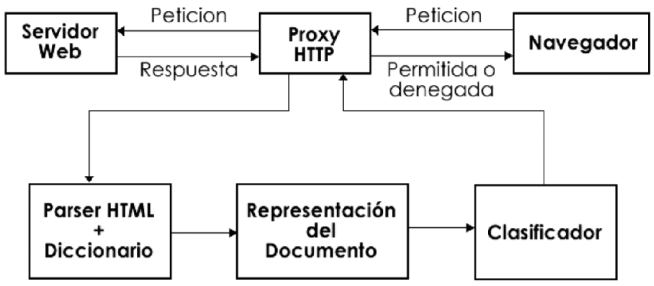


Figura 1. Arquitectura del Software de Filtrado

Clasificador: este componente se encarga de recibir una instancia WEKA y clasificarla. También hace uso del API de WEKA y de una implementación propia del algoritmo KNN.

A. Extracción de información

Las técnicas de extracción de la información hacen referencia en primera instancia al método de recolección de los documentos que se quiere evaluar, en este caso páginas web, y en segunda instancia, a la forma de extraer la información de dichos documentos.

La etapa de recolección de los documentos se realizó de forma manual, considerando las deficiencias y desventajas presentadas por otros métodos descritos a continuación. En primer lugar, no se encontró un conjunto de páginas pornográficas y no pornográficas disponible en la web para utilizarlo, si existen conjuntos de páginas ya recolectados, como se puede ver en [8], donde utilizan un conjunto de páginas web pornográficas y no pornográficas que hacen parte del proyecto POESIA (Public Open-Source Environment for a Safer Internet Access, http://www.poesia-filter.org/), que es un software de código abierto para el filtrado de contenido patrocinado por la Unión Europea. Sin embargo, no se pudo tener acceso a este conjunto de datos porque no esta disponible en la página oficial. Otra opción para esta etapa corresponde a los crawlers que son programas que exploran la web de forma automática y su funcionamiento básico consiste en que a partir de una o un conjunto de URLs extraen los links y los añaden para posteriormente visitarlos y realizar diferentes tareas. Se puede utilizar estas herramientas para que de forma automática descarguen las páginas al disco duro, proporcionándoles un conjunto de URLs base; y que a partir de estas visiten y descarguen las páginas. Sin embargo, la dificultad con este enfoque radica en que no se puede asegurar que si se le proporciona una URL pornográfica todas las páginas que visite van a ser de este tipo; lo cual requiere una validación de cada documento recolectado.

La recolección manual de las páginas aunque supone más trabajo y tiempo, garantiza que la información recolectada corresponde al dominio que se aborda y descarta información ruidosa o que no está dentro del alcance del software (por ejemplo paginas realizadas totalmente en flash). Se recolectaron 2500 páginas (ver Tabla 1).

Tabla 1. Distribución para la recolección de páginas

Tipo	Cantidad	Porcentaje
Pornográficas	750	30%
No pornográficas	1750	70%
Total	2500	100%

Para el proceso de extracción de información de las páginas web es necesario el uso de herramientas denominadas analizadores sintácticos (en inglés parsers). Un analizador sintáctico es una rutina que transforma un texto de entrada en

otras estructuras internas (comúnmente árboles) que facilitan el análisis y capturan la jerarquía implícita de la entrada. En este caso, un analizador sintáctico de HTML captura la estructura del DOM (Document Object Model, http://www.w3.org/DOM/) en un árbol que se puede recorrer para extraer la información pertinente. Para este caso se utilizó el API Jericho 2.6, cuya implementación supone la estructura de un archivo de texto plano en el cual se facilita y agiliza el proceso de extracción.

Existen otras opciones como la librería HTML Parser o NekoHTML que proporcionan funcionalidades similares a Jericho. Sin embargo, se utilizó Jericho debido al rendimiento mostrado comparándola con las otras librerías. El rendimiento superior se debe a las características que lo diferencian de las otras, como por ejemplo que no es un analizador basado en estructuras de árbol sino una combinación de búsqueda de texto y reconocimiento de tags. Además, comparándola con los analizadores basados en árbol los requerimientos de memoria y recursos necesarios para pequeñas secciones del documento son mejores, ignora fácilmente HTML incorrecto o mal formado, y provee una interfaz de mayor nivel e intuitiva para el programador.

B. Diccionario de términos

La clasificación de hipertexto se puede abordar como un problema de clasificación de texto, como se describe en [5]. En este trabajo se eliminan todas las etiquetas HTML del documento a clasificar y se utilizan técnicas propias de Minería de Textos y Recuperación de Información para la clasificación de páginas nocivas (pornografía, terrorismo, drogas, etc.). Una introducción a estas técnicas se puede encontrar en [16] capítulo 1 y 2.

Abordar la clasificación de hipertexto como un problema de clasificación de texto involucra perder el poco conocimiento previo que proporciona la estructura DOM implícita en cada documento. Además, las técnicas de representación de documentos de texto basadas en bolsa de palabras (bag of words), secuencia de palabras, palabras claves, entre otras, pueden presentar ambigüedades y utilizar información no relevante; por ejemplo, el termino cáncer en un documento puede hacer referencia a la enfermedad o al signo zodiacal, lo que representaría una ambigüedad teniendo en cuenta el dominio del problema. De igual forma, la gran mayoría de términos en los documentos de texto no sobrepasan frecuencias mayores a dos, por lo que términos con una sola frecuencia pueden agregar ruido e incrementar el espacio de representación. Es por esto que técnicas basadas en diccionarios pueden mejorar en parte las dificultades de estos enfoques. El diccionario, al construirse enfocándose en un dominio específico contendrá solo términos relevantes a este, lo que ayuda a reducir el ruido como los stopwords (Son palabras que se filtran al principio o al final del procesamiento de texto en áreas como el lenguaje natural o recuperación de información. Generalmente son preposiciones o artículos) y términos que no aportan información a la representación. Con este enfoque se puede realizar una representación de documentos teniendo en cuenta su lenguaje, ya que se construye un diccionario para cada uno, lo que permite que el proceso sea más directo con respecto a técnicas tradicionales de Minería de Textos y Recuperación de Información.

Los métodos basados en diccionario por lo general requieren un conocimiento en el dominio por parte de un experto que se encarga de definirlo. Sin embargo, como no se cuenta con un experto en el dominio de clasificación de páginas pornográficas (Editor Web) que defina un diccionario para este caso, se propone el siguiente enfoque para construirlo: - se recolectan las palabras clave (metatag keywords) de todas las páginas pornográficas y se almacenan en una base de datos, - se eliminan los stopwords encontrados en una lista predefinida, - se realiza un conteo de la frecuencia de las palabras y se ordenan de mayor a menor frecuencia, y - se seleccionan las 150 primeras palabras y éstas constituyen el diccionario.

Se seleccionó la opción iv) debido a que con éste tamaño del diccionario las últimas palabras tienen una frecuencia de aproximadamente 9 o 10, lo que las hace considerables frente a la frecuencia máxima que es de 810, correspondiente a la palabra “porn”. Además, a partir de la palabra 150 sen encuentran palabras muy comunes como “web” y “funny”. No obstante, se puede parametrizar el número de palabras del diccionario a través del proceso de generación del mismo con el fin de realizar otro tipo de pruebas.

C. Representación de las páginas Web

Dentro de este proceso se crea una representación de las páginas la cual corresponde al conjunto de características que permiten describir de la mejor forma su contenido, con el fin de que el modelo de clasificación tenga altos niveles de precisión y un mejor rendimiento. La representación debe contemplar el tipo de páginas que se está abordando y las características del contenido que presentan. En general, la representación y análisis de hipertexto puede clasificarse en dos tipos: basadas en el contenido y basada en los hipervínculos [3], que están directamente relacionados con la Minería web de Contenido y Minería web de Estructura.

Representación basada en el contenido: la Minería web de contenido extrae información semántica de las páginas web. El contenido corresponde a la colección de hechos utilizados en las páginas para brindar la información a los usuarios, es decir, transformar los datos web en conocimiento web. Por ejemplo, comprende texto, imágenes, audio, vídeo o registros estructurados. El contenido HTML incluido en la página objetivo proporciona información útil. La URL en si misma, etiquetas del DOM como el título, los subtítulos y metadatos incrustados en las páginas; como las palabras clave, el lenguaje, etc., permiten describir el contenido de una página web.

Representación basada en los hipervínculos: la web es una amplia colección de documentos unidos entre si mediante

Revista Avances en Sistemas e Informática, Vol.8 No.1, marzo de 2011 - Medellín. ISSN 1657-7663

enlaces o referencias. El lenguaje de comunicación utilizado por cada documento está basado en hipertexto, embebido en código HTML. Éste lenguaje describe la forma en la que debe ser mostrada una página web en un navegador.

De forma general, la web puede ser vista como un grafo dirigido, en el que los nodos son las páginas web y los hipervínculos son representados por URL's. La importancia de esta estructura o topología, se ve reflejada en las tareas desarrolladas por los buscadores web para determinar el ranking o relevancia de cada página. Esta tarea se desarrolla a menudo explorando la referencia a la página en otros documentos en términos de enlaces o de acuerdo al peso o participación de palabras dentro del documento.

Esta aproximación ha tomado gran interés en los últimos años y ha sido objeto de múltiples estudios [4], [7] y [19]. En este enfoque, los hipervínculos de una página y su estructura o topología son estudiados con el fin de extraer información que permita una mejor representación de la página objetivo; información de las páginas hijas, páginas padre, páginas hermanas e información como el texto de anclaje (anchor text) en los hipervínculos son utilizadas para este fin. Además, algoritmos como el Page Rank [1] y HITS [14] son utilizados para determinar la relevancia y autoridad de una página web basándose en la estructura de los enlaces.

Considerando este tipo de representaciones, se decide utilizar una combinación de ambas técnicas y la creación de un diccionario de términos para comparar la información extraída de cada página y proporcionar así mayor efectividad en la representación al igual que se utiliza en [3] y [9]. Los atributos seleccionados dentro de la página son los siguientes:

Características basadas en el contenido:

Titulo(p): número de términos encontrados en el diccionario y presentes en el título de la página p.

Keywords(p): número de términos encontrados en el diccionario y presentes en el metatag keyword de la página p.

Descripción(p): número de términos encontrados en el diccionario y presentes en el metatag descripción de la página p.

TotalPalabras(p): número total de palabras de la página p.

H1(p): número de subtítulos (tag <h1>) de la página p que contienen términos del diccionario.

ImagenesTotales(p): número total de imágenes de la pagina p.

AltImagenes(p): número de imágenes de la pagina p que contienen términos del diccionario en el atributo “alt” o en el atributo “title”.

Características basadas en los hipervínculos

LinksTotales(p): número total de links de la página p.

LinksDiccionario(p): número de links de la página p que contienen en su texto de anclaje (anchor text) términos del diccionario.

TituloEnlaces(p): promedio (número de términos encontrados en el diccionario y presentes en el titulo de las paginas hija (q) para 5 paginas hijas (q) aleatorias de la página p.

KeywordsEnlaces(p): promedio (número de términos encontrados en el diccionario y presentes en los keywords (metatag) de las páginas hijas (q) para 5 paginas hijas (q) aleatorias de la página p.

Estas once características permiten representar la página utilizando los dos enfoques descritos anteriormente. Un punto importante a tener en cuenta son las cinco páginas aleatorias utilizadas en dos de las características basadas en hipervínculos. Se decidió utilizar cinco páginas aleatorias ya que si se utilizan todas las páginas hijas de una página objetivo el costo computacional para realizar la extracción es alto, aún con cinco páginas se puede ver que el tiempo que toma el software para realizar la representación incurre en latencias considerables al usuario final; sin embargo, este valor se dejó parametrizable en el software.

III. CLASIFICACIÓN DE HIPERTEXTO

En Minería de Datos un clasificador primero recibe datos de entrenamiento en los cuales cada entrada es marcada con una etiqueta o clase de un conjunto finito. El clasificador es entrenado usando esos datos, y una vez entrenado, se le proporciona entradas sin etiqueta para que se la asigne [2]. Este mismo procedimiento puede seguirse con documentos de hipertexto, en el cual un clasificador aprende de documentos previamente etiquetados y con esto está en la capacidad de asignarles la etiqueta a nuevos documentos sin clasificar. Las técnicas utilizadas para clasificación de hipertexto puede ser cualquiera utilizada en Minería de Datos, ya que los documentos de hipertexto pueden representarse (ver numeral 2) de tal forma que se ajusten a los requerimientos específicos de cada una.

La selección del algoritmo KNN para la clasificación de las páginas se realizó con base a los resultados de las pruebas e implementaciones con otras técnicas para clasificación de hipertexto: Árboles de decisión C4.5 y Naïve Bayes, también utilizadas en este campo como se puede ver en [22] y [6] [20] respectivamente. La comparación se detalla en el siguiente apartado.

KNN (K-Nearest Neighbors)

KNN ha sido ampliamente utilizado como un efectivo modelo de clasificación [15]. Está basado en una función de distancia que calcula la diferencia o similitud entre instancias [13]. Dada una instancia x, encierra sus k vecinos más cercanos: (y1, y2, …, yk) para asignarle la clase más común denotada por c(x) y determinada por la siguiente ecuación 1:

$$c(x) = \arg\max_{c \in C} \sum_{i=1}^{k} \partial(c, c(y_i)) \quad (1)$$

Donde $c(y_i)$ es la clase de yi y ∂ es una función en donde $\partial(u,v) = 1$, si u = v. La función de distancia más utilizada es la distancia Euclidiana, que se puede definir de la siguiente manera: la distancia euclidiana en medio de los puntos P=(p1, p2, …, pn) y Q=(q1, q2, …, qn) en un espacio n-dimensional se define la ecuación 2:

$$D(P,Q) = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2} \quad (2)$$

El algoritmo KNN tiene tres propiedades claves [17]. - Es un método de aprendizaje perezoso (lazy), es decir que posterga la decisión de cómo generalizar los datos de entrenamiento hasta que una nueva instancia es observada. - clasifica nuevas instancias analizando instancias similares e ignorando las diferentes. Y - representa las instancias como puntos de valor real en un espacio euclidiano n-dimensional. La complejidad computacional está dada por O(np) donde n es el número de instancias y p es el número de atributos. Un estudio más preciso del algoritmo se puede observar en el capítulo 8 de [17] y en el capítulo 5 de [15]. La implementación en la fase de clasificación de hipertexto hace referencia a la programación del algoritmo de aprendizaje computacional que permite determinar si una página web es pornográfica o no. La implementación del algoritmo hace uso del API de WEKA con el fin de reutilizar las estructuras de datos definidas por este y el diseño base de los algoritmos de clasificación. Utilizando esta API, el software de filtrado proporciona flexibilidad ya que en cualquier momento se puede cambiar por otro algoritmo que implemente las mismas interfaces de WEKA.

IV. IMPLEMENTACIÓN Y PRUEBAS

Las pruebas con cada uno de los algoritmos y la configuración básica inicial establecida es: De las 2500 páginas que se obtuvo en la fase de recolección, se dividió el conjunto de datos para la fase de entrenamiento y pruebas. (Ver Tabla 2).

Tabla 2. Total de páginas por tipo para entrenamiento y prueba

Tipo	Entrenamiento	Prueba	Total
Pornográficas	525	225	750 (30%)
No Pornográficas	1225	525	1750 (70%)
Total	1750 (70%)	750 (30%)	2500 (100%)

El total de páginas de entrenamiento corresponde al 70% mientras que el total de páginas de prueba es de 30%. Así mismo, se realizó un proceso de estratificación para que exista el mismo porcentaje de cada una de las clases tanto en el conjunto de entrenamiento como en el conjunto de prueba. El porcentaje de páginas pornográficas en cada uno de los conjuntos corresponde al 30%, mientras que el porcentaje de páginas no pornográficas es del restante 70%.

Para cada algoritmo (Árboles de decisión, Naïve Bayes, y KNN) se realizó pruebas de validación cruzada y de entrenamiento/prueba con la herramienta WEKA con diferentes configuraciones y se concluyó que la configuración por defecto presenta porcentajes de precisión más altos que las otras configuraciones, (aunque estas mejoras no son tan significativas, menores al 1%) por lo que se utilizó dicha configuración.

Posteriormente, al realizar las pruebas en cada uno de los algoritmos se obtuvo el porcentaje general de precisión y las medidas de evaluación clásicas utilizadas en procesos de clasificación, como la precisión, la cobertura y la medida-F (media armónica de la precisión y la cobertura). Para información más de los métodos de evaluación de entrenamiento/prueba y validación cruzada, así como las medidas de evaluación, consultar [17] [15] y [23].

Resultados obtenidos

Los resultados de precisión general de cada uno de los algoritmos con los diferentes métodos de evaluación se pueden observar en la tabla 3. Los resultados de las medidas de evaluación se pueden observar en la tabla 4 y 5.

Tabla 3. Precisión general de los algoritmos de clasificación de hipertexto

Método de Evaluación	Algoritmo	Porcentaje de Precisión
Entrenamiento/prueba	Árbol de decisión C4.5	93.86%
Entrenamiento/prueba	Naive Bayes	94.53%
Entrenamiento/prueba	KNN	94.4%
Validación cruzada (n=10)	Árbol de decisión C4.5	96.8%
Validación cruzada (n=10)	Naive Bayes	96.46%
Validación cruzada (n=10)	KNN	96.85%

Análisis de Resultados

Se evaluaron los resultados obtenidos de cada una de las pruebas realizadas y se eligió un clasificador para usar por defecto en la implementación del filtro.

Tabla 4. Medidas de evaluación con el método Entrenamiento / Prueba

	Pornográficas			No pornográficas		
Algoritmo	precisión	Cobertura	medida-F	precisión	cobertura	medida-F
Árbol de Decisión	0.939	0.84	0.892	0.935	0.981	0.96
Naive Bayes	0.946	0.867	0.905	0.945	0.979	0.962
KNN	0.955	0.853	0.901	0.94	0.983	0.961

Revista Avances en Sistemas e Informática, Vol.8 No.1, marzo de 2011 - Medellín. ISSN 1657-7663

Tabla 5. Medidas de evaluación con el método de validación cruzada

	Pornográficas			No pornográficas		
Algoritmo	precisión	Cobertura	medida-F	precisión	cobertura	medida-F
Árbol de Decisión	0.948	0.945	0.947	0.976	0.978	0.977
Naive Bayes	0.944	0.937	0.941	0.973	0.976	0.975
KNN	0.959	0.935	0.947	0.973	0.983	0.978

WEKA proporciona diversos métodos de selección de atributos o características más relevantes para observar las diferencias existentes entre los tipos de páginas utilizadas y su participación en la clasificación. Se utilizó tres de estos para verificar los atributos que más información aportan al proceso de clasificación. Los tres métodos emplean diferentes medidas de evaluación de los atributos, como la distribución chi-cuadrado (chi-square), ganancia de información (info gain) y la relación de ganancia (gain ratio). La Tabla 6 muestra el orden de relevancia de mayor a menor de cada uno de los atributos con los métodos utilizados, seleccionando como estrategia de búsqueda "Ranker" . Para mayor información y un ejemplo de selección de atributos consultar [10] capitulo 5. Los tres métodos elaboran un escalafón de los atributos más relevantes a los atributos menos relevantes. Se ve que por ejemplo que los atributos relacionados con las páginas hijas (TituloEnlaces(p), KeywordsEnlaces(p)) siempre están en por lo menos los cuatro primeros del escalafón en los tres métodos. Por otra parte, los dos más bajos en el escalafón por los tres métodos son los atributos H1(p) y TotalPalabras(p). En cuanto a clasificación y de acuerdo a los porcentajes de precisión mostrados anteriormente, las mejores técnicas son el Naïve Bayes y KNN (aunque con diferencias mínimas) en cada uno de los métodos de evaluación utilizados. Por otra parte, el costo en tiempo de construcción de los Árboles de Decisión no es tan alto, no obstante, comparándolo con las otras dos técnicas es mayor el tiempo que necesita debido en parte a que las otras dos son técnicas simples en su algoritmia, y no requieren mayores recursos de procesamiento o memoria. Además, la complejidad computacional es mayor en los Árboles de Decisión O(nm2), de tipo polinomial; mientras que la complejidad en las técnicas de Naïve Bayes y KNN O(mnt) y O(np) respectivamente, es de tipo lineal; lo que las hace menos complejas comparándolas con los Árboles de Decisión. De acuerdo a esto, los Árboles de Decisión quedan descartados en primera instancia.

Considerando Naïve Bayes y KNN y teniendo en cuenta que un modelo ideal sería aquel que no filtre o rechace el 100% de contenido no pornográfico y que además filtre el mayor porcentaje del contenido que si lo es; se puede decir que a mayor cobertura de las páginas no pornográficas y mayor precisión en las páginas pornográficas se logrará un mejor modelo. Por lo tanto, y analizando estas medidas, la técnica con mayor nivel de precisión en cada una de estas es KNN por lo que es la técnica seleccionada para el filtro. El aspecto final a evaluar fue las latencias percibidas durante el funcionamiento del filtro considerando que por cada página se extrae información de cinco páginas asociadas. Esto genera retrasos y latencias en la respuesta final al usuario, y se ve en mayor o menor medida dependiendo de la velocidad de conexión a Internet con la que se cuente. Una posible solución a este problema es tener todas las páginas en el disco duro o algún sistema de almacenamiento masivo, lo cual requeriría de una infraestructura robusta y con altos recursos de almacenamiento y cómputo. Una solución de este tipo sería similar a lo descrito en [1].

Tabla 6. Atributos más relevantes

	Método de Selección de Atributos		
Ranking de los atributos	ChiSquaredAttributeEval	IInfoGainAttributeEval	GainRatioAttributeEval
	LinksDiccionario(p)	LinksDiccionario(p)	KeywordsEnlaces(p)
	TituloEnlaces(p)	TituloEnlaces(p)	AltImagenes(p)
	KeywordsEnlaces(p)	Titulo(p)	TituloEnlaces(p)
	Titulo(p)	KeywordsEnlaces(p)	LinksDiccionario(p)
	AltImagenes(p)	AltImagenes(p)	Titulo(p)
	Keywords(p)	LinksTotales(p)	Keywords(p)
	LinksTotales(p)	Keywords(p)	LinksTotales(p)
	ImagenesTotales(p)	Descripcion(p)	ImagenesTotales(p)
	Descripcion(p)	ImagenesTotales(p)	Descripcion(p)
	H1(p)	H1(p)	H1(p)
	TotalPalabras(p)	TotalPalabras(p)	TotalPalabras(p)

V. CONCLUSIONES

Actualmente, el filtrado de contenido web cuenta con herramientas y métodos que requieren gran cantidad de esfuerzo y administración por parte de los usuarios o administradores de red, y su eficiencia se basa en la actualización constante de la información que utilizan para realizar el filtrado. Con el crecimiento constante y acelerado de la World Wide Web, el funcionamiento y diseño de estas herramientas no proporcionan toda la eficacia que se requiere.

El filtrado de contenido web puede abordarse como un problema de clasificación de hipertexto, teniendo en cuenta todas las tareas inherentes como la extracción y la representación de información.

En este artículo se desarrollo un software que utiliza técnicas de Minería web y aprendizaje computacional con el fin de demostrar la aplicabilidad y eficiencia de este enfoque en el filtrado de páginas pornográficas. Se comprobó que los resultados en cuanto a precisión en el filtrado de este tipo de páginas son satisfactorios; por encima del 90% en todas las técnicas analizadas. Gran parte de estos resultados se debe al enfoque

1. Realiza un escalafón de todos los atributos seleccionados utilizando la evaluación individual de cada uno.

que se utilizó para extraer y representar la información; al utilizar en conjunto Minería web de Contenido y Minería web de estructura, se logra recolectar información valiosa que utiliza el clasificador para aprender y clasificar nuevas entradas. Por otra parte, se puede demostrar que el algoritmo KNN presenta algunas ventajas, aunque no tan notorias con respecto a las otras técnicas analizadas.

Un software de este tipo tiene como ventaja analizar la información y tomar la decisión de restringir o no una página en línea, por lo que no depende de información recolectada anteriormente y que requiera constante actualización. El empleo de técnicas de aprendizaje incremental permite al software auto-adaptarse con el fin de responder y actualizar sus mecanismos de decisión. Sin embargo, una desventaja con este enfoque son las altas latencias que percibe el usuario final, debido a los mecanismos inherentes a las técnicas utilizadas.

Este trabajo sirve como aporte al estado del arte en este campo y como punto de partida para la extensión y aplicabilidad de este software, por ejemplo al filtrado de páginas de otros dominios y en diferentes idiomas, la construcción de directorios Web o a la optimización de las búsquedas de información.

VI. TRABAJOS FUTUROS

Estudiar y comparar otras técnicas de representación y clasificación, como por ejemplo enfoques avanzados en Minería de Estructura y algoritmos de aprendizaje como maquinas de vectores de soporte, redes neuronales y en general cualquier técnica que se pueda adaptar al problema. Esto con el fin de identificar y converger en una solución que permita en determinado momento implementar este tipo de técnicas en un ambiente real.

REFERENCIAS

[1] Brin S. y Page L., 1998. The Anatomy of a Large-Scale Hypertextual Web Search Engine. Stanford University. Disponible en http://infolab.stanford.edu/~backrub/google.html.

[2] Chakrabarti S., 2003. Mining de Web: Discovery Knowledge from Hypertext Data. USA: Morgan Kaufmann, 2003. pp. 125-173.

[3] Chau M. y Chen H., 2008. A machine learning approach to web filtering using content and structured analisys, Decision Support Systems. Vol. 44, Issue 2, pp. 482-494.

[4] Cohen W., 2002. Improving A Page Classifier with Anchor Extraction and Link Analysis. In Advances in Neural Information Processing Systems. Vol. 15, pp. 1481-1488.

[5] Fiala D.; Tesař R.; Ježek K. y Rousselot F., 2006. Extracting Information from Web Content and Structure. In Proc. 9th Int. Conf. on Information Systems Implementation and Modelling ISIM'06, Přerov, Czech Republic, pp. 133-140.

[6] Fresno V.; Martinez R.; Montalvo S. y Casillas A., 2006. Naive Bayes Web Page Classification with HTML Mark-Up Enrichment. Proceedings of the International Multi-Conference on Computing in the Global Information Technology, pp. 48.

[7] Glover E.; Tsioutsiouliklis E. K.; Lawrence S.; Pennock D. y Flake G., 2002. Using Web Structure for Classifying and Describing Web Pages. In Proceedings of the eleventh international conference on World Wide Web. Honolulu, Hawaii. Disponible en http://dpennock.com/papers/glover-www-2002-using-web-structure.pdf.

[8] Gomez J.; Carrero F. y Puertas E., 2005. Named Entity Recognition for Web Content Filtering. Natural Language Processing and Information Systems, pp. 286-297.

[9] Gy S.; JH L.; YH M. y SH L., 2004. Improving the precision of the keyword-matching pornographic text filtering method using a hybrid model. Journal of Zhejiang University Science. Vol. 5 No.9, pp. 1106-13.

[10] Hernández J.; Ramírez M. y Ferri C., 2004. Introducción a la Minería de Datos, España: Prentice Hall. pp. 97-125.

[11] Internet World Stats, Internet Usage and Population in South America. [Online], diciembre de 2009, Disponible: http://www.internetworldstats.com/stats15.htm

[12] Internet World Stats. Internet World Users by Language, diciembre de 2009, Disponible: http://www.internetworldstats.com/stats7.htm

[13] Jiang L.; Zhang H. y Su J., 2005. Learning k-Nearest Neighbor Naive Bayes For Ranking Advanced data mining and applications, Vol. 3584, pp. 175-185.

[14] Kleinberg J., 1999. Authoritative sources in a hyperlinked environment. Journal of the ACM (JACM). Vol. 46, pp. 604-632.

[15] Larose D., 2004. Discovering Knowledge in Data. An Introduction to Data Mining, USA: Wiley-Interscience. pp. 90-106.

[16] Markov Z. y Larose D., 2007. Data Mining the Web. Uncovering Patterns in Web Content, Structure, and Usage, USA: Jhon Wiley & Sons. pp. 3-57.

[17] Mitchell T., 1997. Machine Learning. USA: McGraw-Hill. pp. 230-247.

[18] Netcraf, Web Server Survey, junio de 2009, Disponible: http://news.netcraft.com/archives/2009/06/17/june_2009_web_server_survey.html

[19] Prakash A. y Kumar K., 2001. Web Page Classification based on Document Structure. International Institute of Information Technology. Hyderabad, India. Disponible en http://www.iiit.net/students/stud_pdfs/kranthi1.pdf.

[20] Riboni D., 2002. Feature Selection for Web Page Classification. In A min Tjoa, A. C. S., editor, EURASIAICT2002 Proceedings of the Workshops.

[21] Ropelato. J. Internet Pornography Statistics, marzo de 2009, Disponible:
http://internet-filter-review.toptenreviews.com/internet-pornography-statistics-pg4.html

[22] Tsukada M.; Washio T. y Motoda H., 2001. Automatic Web Page Classification by Using Machine Learning. WI: web intelligence: research and development. Disponible en http://www.ar.sanken.osaka-u.ac.jp/papers/2006-12/wi01_tsukada.pdf, pp. 303-313.

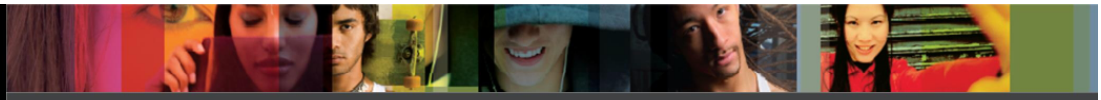
[23] Witten I. y Frank E., 2005. Data Mining, Practical Machine Learning Tools and Techniques, Second Edition, USA: Morgan Kaufmann. pp. 143-184.

Jorge Rodríguez. Magister en Ingeniería de Sistemas. Especialista en Diseño y Construcción de Soluciones Telemáticas. Especialista en Ingeniería de Software. Ingeniero de Sistemas. Docente investigador de la Universidad Distrital "Francisco José de Caldas".

Harry Barrera. Especialista en Ingeniería de Software. Ingeniero en Telemática y Tecnólogo en Sistematización de Datos de la Universidad Distrital "Francisco José de Caldas".

Sandra Bautista. Ingeniera en Telemática y Tecnóloga en Sistematización de Datos de la Universidad Distrital "Francisco José de Caldas"

Revista Avances en Sistemas e Informática, Vol.8 No.1, marzo de 2011 - Medellín. ISSN 1657-7663



UNIVERSIDAD NACIONAL DE COLOMBIA SEDE MEDELLÍN

Cisco | Networking Academy®
Mind Wide Open™

Cisco Networking Academy es un programa ampliamente conocido de e-doing que enseña a los estudiantes las habilidades tecnológicas de Internet en una economía global. El programa proporciona contenido basado en la Web, pruebas en línea, seguimiento del desempeño de los estudiantes, laboratorios con equipos reales y con simuladores, soporte y entrenamiento por parte de los instructores, así como preparación para las certificaciones estándares de la industria.

Oferta de cursos

- Mantenimiento de PC: IT Essentials
- Redes básicas: Cisco Certified Network Associate
- Redes avanzadas: Cisco Certified Network Professional
- Seguridad en routers: CCNA Security
- Voz sobre IP
- Asterisk básico

Programación 2011

Ciclo 48: Inicia 17 de enero. Finaliza 12 de marzo
Ciclo 49: Inicia 22 de marzo. Finaliza 23 de mayo
Ciclo 50: Inicia 30 de mayo. Finaliza 29 de julio
Ciclo 51: Inicia 8 de agosto. Finaliza 3 de octubre
Ciclo 52: Inicia 10 de octubre. Finaliza 12 de diciembre

Consulte los horarios de cada nivel a través de nuestros canales informativos al pie de página

Además...

- Alquiler de laboratorios virtuales para auto-estudio o cursos empresariales
- Presentación de exámenes de certificación para multiples areas bajo el Centro Pearson VUE
- Cursos exclusivos para su empresa
- Pregunte por nuestros descuentos

CATC - Academia Regional - Academia Local
Universidad Nacional de Colombia sede Medellín
Calle 65 78-28 Bloque M1 Oficina 101. Facultad de Minas
Teléfono: +57 4 4255268 Fax: +57 4 2341002 E-mail: **catc@unal.edu.co** Web: **http://cnap.unalmed.edu.co**
Facebook: **fb.me/catcunal** Twitter: **@catcunal** Buzz: **google.com/profiles/catcunal**
Medellín, Colombia

cisco. Networking Academy