Revista Cubana de Ciencias Informáticas
ISSN: 1994-1536 | e-ISSN: 2227-1899
RNPS: 0547 | e-RNPS: 2301
http://rcci.uci.cu

Tipo de artículo: artículo original
Temática: Inteligencia artificial
Recibido: 10/12/2011 | Aceptado: 24/1/2012

Técnicas de Inteligencia Artificial en el filtro de contenido web Smart Keeper para la clasificación de información.

Artificial Intelligence techniques in the Smart Keeper web content filter for classification of information.

Yurisleidy Hernández Moya[1*], Dovier Antonio Ripoll Méndez[2], Luis Enrique Sánchez Arce[3], Kiuver Kaddiel Ibañez Castro[4], Karel Antonio Verdecia Ortiz[5]

1* Departamento Soluciones Informáticas para Internet. Centro de Ideoinformática. Universidad de las Ciencias Informáticas. Carretera a San Antonio de los Baños, km 2 1⁄2, Torrens, Boyeros, La Habana, Cuba, CP 19370.
ymoya@uci.cu
2 Departamento Docente Central Técnicas de Programación. Universidad de las Ciencias Informáticas. Carretera a San Antonio de los Baños, km 2 1⁄2, Torrens, Boyeros, La Habana, Cuba, CP. 19370.
daripoll@uci.cu
3 Departamento Soluciones Informáticas para Internet. Centro de Ideoinformática. Universidad de las Ciencias Informáticas. Carretera a San Antonio de los Baños, km 2 1⁄2, Torrens, Boyeros, La Habana, Cuba, CP 19370.
lesanchez@uci.cu
4 Departamento Soluciones Informáticas para Internet. Centro de Ideoinformática. Universidad de las Ciencias Informáticas. Carretera a San Antonio de los Baños, km 2 1⁄2, Torrens, Boyeros, La Habana, Cuba, CP 19370.
kiuver@uci.cu
5 Departamento Soluciones Informáticas para Internet. Centro de Ideoinformática. Universidad de las Ciencias Informáticas. Carretera a San Antonio de los Baños, km 2 1⁄2, Torrens, Boyeros, La Habana, Cuba, CP 19370.
kverdecia@uci.cu

Resumen: En los filtros de contenido web resulta importante identificar la temática de la información a la que acceden los usuarios, con el fin de ser capaz de decidir si permitirla o denegarla. El análisis inteligente de contenido es una de las técnicas empleadas por este tipo de software para la clasificación de información. Para el filtro cubano de contenido Smart Keeper se desarrolla el software Motor de Clasificación Inteligente de Contenido. El motor emplea técnicas de Inteligencia Artificial con el propósito de clasificar las páginas web. Se realizaron pruebas a componentes independientes del motor y los resultados obtenidos han sido favorables. En este trabajo se abordan algunas técnicas de Inteligencia Artificial empleadas en Smart Keeper durante el proceso de clasificación de la información.

Palabras clave: Filtro de contenido, Inteligencia Artificial, Smart Keeper, Motor de Clasificación Inteligente de Contenido.

***Abstract**: In web content filters is important to identify the subject of the information accessed by users to be able to decide whether to allow or deny. Intelligent Content Analysis is one of the techniques employed by this type of software for the classification of information. The Engine of Intelligent Content Classification software being developed to be used with the cuban content filtering system Smart Keeper . The engine use Artificial Intelligence techniques in order to classify web pages. Tests were conducted to separate components of the engine and the results have been favorable. In this paper we address some techniques of Artificial Intelligence used in Smart Keeper during the process of classifying information.*
***Keywords**: Content filter, Artificial Intelligence, Smart Keeper, Engine of Content Intelligent Classification.*

1. Introducción

En Internet se encuentran grandes volúmenes de contenidos de diversas fuentes y temáticas que pueden variar desde materiales educativos, de ocio y científico-técnico hasta los considerados inadecuados (nocivos e ilícitos) (García, 2005). La información perteneciente a tales clasificaciones puede variar de un país a otro e incluso dentro de un mismo territorio. Diversos ejemplos evidencian lo anterior: en Alemania, Hungría, Francia y Austria existen leyes que prohíben la exhibición de símbolos nazis lo cual resulta legal en los Estados Unidos (Antisemitismo, 2010) o en países como Holanda donde es permitido ejercer la prostitución no siendo así en otros como Cuba (Trejo y Álvarez, 2007). Por tal motivo resulta difícil establecer una norma global para determinar qué elementos son permitidos o no publicar y/o ver en Internet.

Varios países han establecido leyes para regular la navegación de sus usuarios como la Ley de Decencia de las Comunicaciones de Estados Unidos, la cual finalmente fue rechazada por atentar contra la libertad de expresión (Sánchez, 2004). En Argentina la ley 25.690 indica a los Proveedores de Servicio de Internet la obligación de ofrecer software de protección que impida el acceso a sitios específicos (Torres, 2003).

Por otra parte, comúnmente instituciones que ofrecen servicio de Internet establecen su propia Política de Uso Aceptable de Internet, constituida por un conjunto de reglas que restringen las formas en que los sistemas informáticos pueden ser usados (Fernández, 1994), en correspondencia con la función que realizan. En ocasiones, el

establecimiento de estas regulaciones no significa medida suficiente para su cumplimiento. En este sentido, los filtros de contenido emergen como una solución informática que contribuyen al cumplimiento de las normas establecidas de navegación.

Generalmente instituciones cubanas con acceso a Internet, establecen normas para regular dicho servicio. Específicamente la Universidad de las Ciencias Informáticas (UCI) expresa en su reglamento, entre otros elementos, la prohibición de transmitir, acceder o difundir información pornográfica, terrorista, contrarrevolucionaria, religiosa u otros fuera de los intereses de la institución mediante cualquier servicio de la red incluyendo Internet (Gil, 2008). Acorde con su misión la UCI desarrolla el filtro de contenido web denominado Smart Keeper, con el fin de apoyar a las instituciones en el control de los materiales que acceden sus miembros en Internet.

Los filtros de contenido son elementos de software que permiten o deniegan el acceso a materiales basados en el contenido. Filtrado de virus, correos electrónicos, páginas web, mensajería instantánea, entre otros, son algunas de las funcionalidades que podrían poseer estos sistemas. En el presente trabajo se abordará únicamente lo relacionado con el filtrado de contenido web.

Cuando una página web es solicitada, el filtro de contenido debe ser capaz de decidir si el material es mostrado o no al usuario. Algunas de las técnicas empleadas por los filtros para el análisis de la información son: la Plataforma para la Selección de Contenido en Internet, listas de sitios considerados inaceptables (listas negras), búsquedas de palabras clave y análisis inteligente de contenido (Lee, 2002). Filtros comerciales generalmente emplean el análisis inteligente de contenido y las listas de URLs (Lin et al, 2008).

Hasta el momento Smart Keeper emplea listas de URLs categorizadas por terceros disponibles en Internet. Con el objetivo de eliminar la dependencia de tales listas se desea realizar el análisis inteligente de la información. Por tal motivo resulta de interés identificar técnicas para categorizar componentes de una página web tales como: el texto, las imágenes y los enlaces.

2. Materiales y Métodos

Varios filtros de contenidos afirman el uso de algoritmos de Inteligencia Artificial en su solución aunque algunos no ofrecen mayores detalles debido a que constituyen un secreto comercial. No obstante resulta oportuno analizar algunos de estos:

- POESIA (Public Open source Environment for Safer Internet Access) es un software de código abierto financiado por la Unión Europea. El filtro posee varios componentes, algunos de los cuales son encargados de analizar imágenes, reconocer idiomas y categorizar textos. Con el resultado de estos componentes el sistema determina si la página debe permitirse o no. En la identificación del idioma es empleado el método supervisado Out-of-Place y para el procesamiento de las imágenes son utilizados algoritmos de detección de piel y formas. El Aprendizaje Estadístico es empleado para el análisis del texto y las Máquinas de Soporte Vectorial para la construcción del clasificador. Con el Modelo del Espacio Vectorial el sistema logra una representación vectorial de los documentos y además emplea algunas técnicas de Procesamiento del Lenguaje Natural para extraer la raíz de las palabras (Puertas et al, 2004).
- Surf-SeCure de la empresa PineApp™: provee un sistema de filtrado que inspecciona los tráficos HTTP y FTP. Aplicaciones como mensajería instantánea, VoIP y juegos en línea son identificadas y bloqueadas por este sistema. Surf-SeCure realiza la clasificación en "tiempo real" de páginas web empleando el Reconocimiento Artificial de Contenido™, un motor basado en tecnología de Inteligencia Artificial de reconocimiento activo de contenidos (PineApp, 2007).
- PureSight: basado en la tecnología Reconocimiento Artificial de Contenido™ (PURESIGHT, 2004). Reconocimiento Artificial de Contenido™ recibe paquete a paquete la página web solicitada por el usuario. Usando un analizador extrae cientos de características de la página HTML y conforma con ello un vector de datos. Mediante un extractor de características, que emplea algoritmos de Inteligencia Artificial, se determinan los rasgos del vector de datos que resultan más útiles en la categorización de la página, creando así un nuevo vector de datos procesados con menos información en comparación con el vector anterior. Un mecanismo de Clustering es empleado para tomar las combinaciones y relaciones de las características del vector de datos procesados, resultando una única coordenada matemática. El Reconocimiento Artificial de Contenido™ clasifica la página cuando ubica la coordenada dentro de una categoría de información previamente definida (Communications, 2009).
- WebFilter™: uno de los productos de la compañía Optenet para regular el acceso a contenido de Internet. Optenet WebFilter™ identifica, clasifica y bloquea el acceso a páginas web inapropiadas. El sistema emplea varias tecnologías, tales como: Optenet GIANT™ (recibe flujos de información de Internet para actualizar las soluciones de la empresa), Optenet CCOTTA™ (analiza la información en "tiempo real", redireccionándola a los servicios adecuados de filtrado) y Optenet MIDAS™ (permite la identificación de contenidos ilegales y nocivos) (Optenet, 2011).

Para el desarrollo de Smart Keeper se seleeciona Symfony con PHP 5 como framework para la creación de la Interfáz de Administración por su compatibilidad con el gestor PostgreSQL, seguridad y automatización de varias tareas comunes en proyectos web. Fueron además utilizados los lenguajes de programación C/C++ y Python pues varias de las bibliotecas empleadas fueron desarrolladas con estos lenguajes.

3. Resultados y discusión

Smart Keeper es un filtro cubano de contenido web orientado a servidores para medianas redes de usuarios como las disponibles generalmente en colegios, empresas y bibliotecas. Dicho filtro es flexible a la política de uso de Internet del lugar donde se aplique y cuenta con una interfaz web fácil e intuitiva para la administración. Mediante la definición de políticas y grupos, posibilita establecer diversos permisos de navegación por tipos de usuarios. Para su funcionamiento el sistema emplea las listas de URLs categorizadas: Isak y Toulouse[1]. Además este filtro cuenta con un componente, aún en desarrollo, denominado Motor de Clasificación Inteligente de Contenido (MOCIC) que, empleando algoritmos de Inteligencia Artificial, será el encargado de clasificar las páginas web.

MOCIC

El motor está integrado por varios módulos que analizan componentes de las páginas como el texto, los enlaces y las imágenes. Finalmente con el criterio obtenido de cada uno de tales componentes, el sistema asigna una o varias categorías a la URL.

Módulo: Identificador de Idioma

MOCIC realiza la identificación de idioma como un paso previo a la categorización de texto, pues en dependencia del idioma empleado se determina qué método debe ser utilizado para catalogarlo. En la tarea de reconocer el idioma se emplean las Palabras Vacías del texto, tales como: artículos, pronombres, preposiciones, entre otras (Sinka y Corne, 2003). Tales palabras ofrecen poca información del tema tratado en el contenido, pero sí resultan de interés para determinar el lenguaje. Actualmente el módulo emplea listas de Palabras Vacías obtenidas de Snowball[2] y es capaz de reconocer seis idiomas: inglés, español, francés, alemán, portugués e italiano.

[1] http://cri.univ-tlse1.fr/blacklists/download/blacklists.tar.gz
[2] http://snowball.tartarus.org/

Módulo: Categorizador de Texto

Para la categorización del texto se emplea la clasificación supervisada, tomando como entrenamiento una colección de documentos previamente categorizada y obtenida de BankSearch DataSet[3]. Tal colección está conformada por aproximadamente 11 000 páginas web y organizada en 11 categorías temáticas: Astronomía, Biología, Cajas de Ahorro, C/C++, Bancos Comerciales, Agencias de Seguros, Java, Motores de Carrera, Visual Basic, Fútbol y Otros Deportes. Para el análisis del texto el Modelo de Espacio Vectorial es utilizado en la representación de los documentos, mientras que para la clasificación puede escogerse entre los siguientes algoritmos de Aprendizaje Automático (AA): Máquina de Soporte Vectorial, K-NN y Rocchio.

La Tabla 1 muestra resultados obtenidos por el módulo empleando el algoritmo Máquina de Soporte Vectorial. Los valores de Precisión y Exhaustividad fueron calculados a partir de las siguientes fórmulas:

$$R = \frac{a}{a+b} \quad (1)$$

$$P = \frac{a}{a+c} \quad (2)$$

donde:

R es la exhaustividad.

P es la precisión.

a es el número de documentos pertenecientes a una clase y asignados a la misma.

b es el número de documentos no pertenecientes a una clase y asignados a la misma.

c es el número de documentos pertenecientes a una clase que no fueron asignados a la misma.

[3] http://www.genalgo.com/index.php?option=com_content&task=view&id=113&Itemid=2

Tabla 1. Resultados obtenidos por el categorizador de textos con el algoritmo Máquina de Soporte Vectorial.

Categoría	Cantidad de Documentos	Clasificados Correctamente	% Clasificados Correctamente	Precisión	Exhaustividad
Astronomía	189	186	98.41	0.9841	0.9841
Biología	180	177	98.33	0.9672	0.9833
Cajas de Ahorro	180	172	95.56	0.9149	0.9556
C/C++	181	170	93.92	0.9497	0.9392
Bancos Comerciales	208	178	85.58	0.9570	0.8558
Agencias de Seguros	189	184	97.35	0.9200	0.9460
Java	198	195	98.48	1.0000	0.9848
Motores de Carrera	193	187	96.89	0.9397	0.9689
Otros Deportes	209	208	99.52	0.9952	0.9952
Fútbol	184	183	99.46	0.9946	0.9946
Visual Basic	180	177	98.33	0.9888	0.9833
Total	2091	2017			

La Tabla 2 muestra la matriz de confusión a partir de la evaluación realizada. Las filas indican la cantidad de elementos de la categoría y las columnas reflejan en qué clase exactamente fueron ubicados tales elementos por el categorizador.

Tabla 2. Matriz de confusión a partir de los resultados obtenidos con el categorizador de textos.

Categoría	AA	AB	AC	AD	AE	AF	AG	AH	AI	AJ	AK
Astronomía (AA)	186	2	0	0	1	0	0	0	0	0	0
Biología (AB)	1	177	0	1	0	0	0	1	0	0	0
Cajas de Ahorro (AC)	0	0	172	0	5	2	0	1	0	0	0
C/C++ (AD)	1	1	1	170	0	0	0	7	0	0	1
Bancos Comerciales (AE)	0	0	15	0	178	14	0	1	0	0	0
Agencias de Seguros (AF)	0	1	0	0	2	184	0	0	1	1	0
Java (AG)	1	0	0	1	0	0	195	0	0	0	1
Motores de Carrera (AH)	0	1	0	5	0	0	0	187	0	0	0
Otros Deportes (AI)	0	0	0	0	0	0	0	1	208	0	0
Fútbol (AJ)	0	0	0	1	0	0	0	0	0	183	0
Visual Basic (AK)	0	1	0	1	0	0	0	1	0	0	177

La categoría Bancos Comerciales muestra los peores resultados; en este sentido se podría modificar la colección de entrenamiento para esta categoría y observar nuevamente la salida del algoritmo. En la categoría de Otros Deportes se obtienen los mejores valores. En sentido general los resultados son favorables lo que resulta alentador para MOCIC.

Módulo: Analizador de Enlaces

Mediante hipervínculos es posible relacionar diversos materiales digitales. Los documentos son enlazados generalmente cuando sus contenidos guardan algún tipo de relación. Por tal motivo, en una página web los enlaces entrantes (que conducen a la página en cuestión) y salientes (que guían desde la página hacia otra) constituyen otro aspecto que podría aportar información sobre el contenido de la misma. Varios trabajos han sido realizados con respecto a este tema. En (Oh, Myaendg y Lee, 2000) emplean las categorías de las páginas inmediatas (las conectadas por hipervínculos) para establecerle una categoría a los nuevos elementos. Por otra parte, (Sen P. et al, 2008) comparan el rendimiento de tres algoritmos en la tarea de clasificación basada en enlaces, a partir de gráficos irregulares posiblemente compuestos por ciclos, determinando el Algoritmo de Clasificación Iterativa como el más confiable.

Con una base de datos de URLs categorizadas, el Analizador de Enlaces de MOCIC utiliza técnicas de Minería de Datos para determinar heurísticas que posibilitan clasificar las páginas web teniendo en cuenta sus enlaces entrantes y salientes.

Módulo: Procesador de Imágenes Digitales

El procesamiento de imágenes realizado en MOCIC está enmarcado en el campo de la visión por computadora. Durante el estudio de las imágenes se realizan diferentes procesos como el reconocimiento de piel, objetos (esvásticas, símbolos religiosos, entre otros), rostros, caracteres e imágenes con personas desnudas. Típicamente se emplean algoritmos de AA que parten de una base de conocimiento.

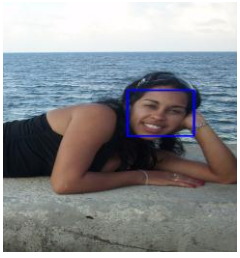
Para determinar en una imagen las regiones que conforman piel se emplean las Redes Neuronales Artificiales. La red es entrenada con varias imágenes segmentadas con anterioridad. La Figura 1a muestra una imagen seleccionada y a continuación (Figura 1b), un mapa de probabilidad de la piel obtenido por el módulo. Tal mapa es una imagen en escala de grises donde cada pixel posee una valor entre 0 y 1 indicando su grado de pertenencia a la clase piel. Con el empleo de un umbral se toma la decisión final (piel o no) por cada pixel; la Figura 1c muestra el resultado de emplear un umbral con valor 0.95.

a) b) c)

Figura 1: Resultado del proceso de segmentación en piel o no piel: a) Imagen seleccionada, b) Mapa de probabilidad de la piel, c) Imagen binaria.

Para el reconocimiento de objetos se usan los métodos Adaboost y Scale-Invariant Feature (SIFT). Este último recibe como entrada una imagen, correspondiente al objeto a buscar. AdaBoost es utilizado específicamente en la identificación de rostros, el cual fue entrenado con aproximadamente 3 000 imágenes de rostros que fueron obtenidas a partir del directorio de personas de la UCI. La Figura 2a muestra el resultado del módulo al identificar rostros en la Figura 1a.

Páginas web sobre determinados temas como pornografía, poseen generalmente fotografías con personas desnudas; por tal motivo identificar este tipo de imágenes es una tarea de interés para MOCIC. El módulo para la detección de imágenes desnudas realiza primero la segmentación en piel o no piel, como fue explicado anteriormente, obteniendo una imagen binaria de la cual se obtienen varios indicadores entre los que se encuentran: cociente entre la cantidad de píxeles con color de la piel y el total de píxeles de la imagen, cantidad de regiones con color de la piel, cantidad de píxeles detectados con el color de la piel en la región más grande y cociente entre la cantidad de píxeles con el color



de la piel en la región más grande y el total de píxeles con color de la piel en la imagen.

Figura 2: Resultado del proceso de identificación de rostros.

Las Redes Neuronales Artificiales, específicamente el Perceptrón Multicapa, es utilizado por este módulo para tomar

la decisión final; el algoritmo es instruido a partir de una base de conocimiento con aproximadamente 300 imágenes desnudas segmentadas con anterioridad.

4. Conclusiones

- MOCIC dotará a Smart Keeper de una herramienta capaz de realizar análisis de contenido mediante técnicas de Inteligencia Artificial.
- Con el funcionamiento del motor el filtro, para la actualización de su base de datos, tendrá la posibilidad de independizarse de las listas de URLs categorizadas por terceras partes.
- El empleo de Minería de Datos y algoritmos de Aprendizaje Automático han mostrador resultados favorables en el análisis de componentes de una página web.

5. Referencia

Antisemitismo, f. *Legislacion contra el antisemitismo y la negacion del holocausto*. [en línea] *The Coordination Forum for Countering Antisemitism* 2010. [Consultado el: 29 Septiembre 2011]. Disponible en :http://www.antisemitism.org.il/spa/legislacion%20contra%20el%20antisemitismo%20y%20la%20negacion%20del%20holocausto

Communications, A. *Advanced Content Recognition (ACR) Technology An Overview of the AI Technology Inside Allot NetPure.* 2009. p: 4

Fernández, F. C. *Glosario básico inglés-español para usuarios de Internet.* 1994. p.49.

García, M. J. *Regulación y autorregulación en Internet: el control de los contenidos y los datos en la LSSI*. 12, [sede de la Agencia Catalana de Protección de Datos], Noviembre 2005.

Gil, M. *Resolución no. 299/09.* 2009. Universidad de las Ciencias Informáticas.

Lee P.; Hui S.; Fong A. *Neural networks for Web content filtering*. IEEE Intelligent Systems. 2002, 17(5), p.48-57

Lin P. *et al*. *Accelerating Web Content Filtering by the Early Decision Algorithm.* IEICE Transactions. 2008, E91-D(2), p. 251-257

Oh H.; Myaendg S.; Lee M. *A practical hypertext categorization method using links and incrementally available class information*. Proceedings of SIGIR 23rd ACM International Conference on Research and Development in Information Retrieval, 2000, ISBN:1-58113-226-3: p. 264-271

Optenet. *Optenet WebFilter*. [en línea]. 2011. [Consultado el: 10 de Obtubre de 2011]; Disponible en: http://www.optenet.com/es/webfilter.asp

PineApp. *Surf-Secure Filtrado Web en tiempo real.* 2007

Puertas, E. *et al*. *Filtrado de contenidos web en español dentro del proyecto poesia*. En: *Conferencia Ibero-Americana WWW/Internet*. 2004. p. 5

PureSight. *PureSight User's Guide*. 2004. p:139

Sánchez, C. *La ley de decencia en las comunicaciones y GILC*. República Internet. Barcelona. Julio 2004, p. 15-16

Sen P. *et al*. *Collective Classification in Network Data*. AI Magazine. 2008, 29 (3): p. 93-106

Sinka M.; Corne. D. *Evolving better stoplists for document clustering and web intelligence*. Proceedings of the 7th WSEAS Int'l Conf. on Artificial Intelligence, 2003, p.1015–1023.

Torres, C. *Ley 25.690*. Enero 2003

Trejo, E.; Álvarez M. *Estudio de Legislación Internacional y Derecho Comparado de la Prostitución*. Junio 2007.