

## HERRAMIENTA PARA EL ANÁLISIS DE LA PERTINENCIA DE LA NAVEGACIÓN POR INTERNET

Ing. Susana Pérez Mantilla<sup>1</sup>, Ing. Aryam Gutiérrez López<sup>2</sup>, Dr. C Alain A. Garófalo Hernández<sup>3</sup>

1 Departamento de Telecomunicaciones y Telemática. Facultad de Ingeniería Eléctrica, CUJAE, susana.pm@electronica.cujae.edu.cu

2 Departamento de Telecomunicaciones y Telemática. Facultad de Ingeniería Eléctrica, CUJAE, aryam@electronica.cujae.edu.cu

3 Departamento de Telecomunicaciones y Telemática. Facultad de Ingeniería Eléctrica, CUJAE, [alain@tesla.cujae.edu.cu](mailto:alain@tesla.cujae.edu.cu)

### RESUMEN

El uso desmedido de Internet en horario laboral para fines personales, es una práctica muy difundida en la mayoría de las empresas. Esta situación provoca que la red esté constantemente cargada, la insatisfacción de los usuarios y la afectación al desempeño de aplicaciones críticas; además del costo monetario que representa pagar por un ancho de banda que no se está utilizando para lo que realmente se contrató. Es por ello que el estudio de las técnicas de clasificación de tráfico a Internet, desde el punto de vista de pertinencia, juega un papel fundamental para realizar un análisis del consumo.

El objetivo fundamental de este trabajo es desarrollar una aplicación que emita un criterio acerca de la pertinencia de la navegación por Internet en redes empresariales. El Sistema de Evaluación de Pertinencia (SEP) es una herramienta que le permite al administrador definir, de acuerdo a los objetivos de la empresa, un patrón contra el cual se van a comparar los accesos a Internet de sus usuarios para evaluarlos en distintos grados de pertinencia.

**PALABRAS CLAVES:** clasificación de páginas web, navegación por Internet, pertinencia

### TOOL TO ANALYZE THE PERTINENCE OF WEB SURFING

### ABSTRACT

Overuse of the Internet during work hours for personal purposes is a widespread practice in most companies. This situation causes that the network is constantly charged, the user dissatisfaction and the affectation of the performance of critical applications, in addition to the monetary cost of paying for bandwidth that is not being used for what is actually hired. That is why the study of techniques for internet traffic classification plays a key role in the analysis of consumption from the point of view of pertinence.

The main objective of this work is to develop an application to issue a criterion about the adequacy of the Internet browsing in enterprise networks. The Pertinence Assessment System (SEP) is a tool that allows the administrator to define, according to the objectives of the company, a standard against which the internet access of users will be compared to assess the different degrees of pertinence.

**KEY WORDS:** web classification, Internet surfing, pertinence

## **INTRODUCCIÓN**

El uso de Internet para fines personales en horario de trabajo es una práctica muy difundida en todo el mundo. Muchas veces los usuarios se quejan en la oficina por lo lenta que está la conexión a Internet. Es bastante común que este problema tenga que ver con el uso indiscriminado que se hace de la web dentro de la empresa. Las empresas tratan de controlarlo cada vez más, porque repercute en sus costos y en el menor tiempo que los empleados trabajan. Esto sin tener en cuenta los riesgos informáticos que implica la navegación por sitios poco seguros. [1]

Es de interés de los altos directivos de las empresas, conocer cuáles trabajadores son más productivos y eficientes en su labor. Para conocer esto se deberían tener en cuenta varios factores, uno muy importante sería el nivel de pertinencia de la navegación por Internet. El nivel de pertinencia no es más que el grado de relación positiva que existe entre las páginas web que se visitan y el objetivo social de la organización. [2]

## **HERRAMIENTAS PARA LA CLASIFICACIÓN DEL TRÁFICO WEB**

Para realizar el control del uso de Internet por parte de las empresas se han desarrollado un gran número de herramientas que permiten conocer con detalle el acceso de cada usuario, como los sistemas de reporte; además de otras que permiten restringir el acceso a determinados sitios, como los filtros de contenido.

Los sistemas de reporte tienen como función principal analizar las bitácoras generadas por los servidores proxies y generar un reporte que le facilite al administrador la interpretación de dichas bitácoras. Algunos de los más conocidos son SRG1 y SARG2 de libre distribución, IAM3 y SurfCop de pago, y TopInt desarrollado en la CUJAE. Los sistemas de reporte, a pesar de conocer cada uno de los accesos a Internet, son incapaces de emitir un juicio automático acerca de cuán pertinente fue el uso que le dieron los usuarios a Internet. [3][4][5]

Por otra parte, los filtros de contenido sí implementan un método de análisis de las páginas que solicitan los usuarios y realizan de alguna manera una clasificación de las mismas. Sin embargo, esta clasificación tiene como objetivo impedir el acceso a determinadas páginas. Una de las técnicas que implementan los filtros es la clasificación por el análisis semántico de la página. Este método consiste en asignarle valores de peso, previamente definidos, a cada una de las palabras de la página y luego mediante una ecuación matemática, obtener un valor de peso o rating de la página. Este peso final es comparado con un valor límite de peso admisible, si lo sobrepasa, la página es bloqueada, si no, el usuario podrá ver la página. Este método de clasificación es utilizado por filtros de contenido como DansGuardian y POESIA. [6][7]

Tanto los sistemas de reporte como los filtros de contenido desempeñan un papel importante a la hora de conocer y controlar el acceso a Internet, pero no son capaces de automáticamente generar valores de pertinencia que reflejen en realidad cómo se están aprovechando los recursos de la red para cumplir con el objetivo social de la empresa. Los filtros de contenido cuentan con la ventaja de poder dar algún tipo de clasificación a las páginas a las que acceden los usuarios y registrar esta información en sus reportes, lo que pudiera darles a los administradores la posibilidad de evaluar el tráfico manualmente. Sin embargo, este trabajo es extremadamente complicado debido a que requiere de mucho tiempo de los administradores para analizar todas las entradas de los registros, y en redes grandes en las que el acceso a Internet es constante se hace imposible. Por otro lado, hacer un análisis a partir de los registros

de los sistemas de reporte, trae consigo la necesidad de una gran preparación y conocimiento de Internet por parte de los administradores, que tendrían, la mayoría de la veces, que acceder al sitio para decidir su grado de pertinencia. [2]

## SISTEMA DE EVALUACIÓN DE PERTINENCIA (SEP)

El Sistema de Evaluación de Pertinencia (SEP) surge con el objetivo principal de clasificar a los usuarios de la red, de acuerdo a sus accesos a Internet, en más o menos pertinente. El resultado de esta clasificación está vinculado en gran medida con los fines y objetivos de la organización, puesto que el administrador definirá parámetros que formarán un patrón contra el cual se compararán los accesos de los usuarios. Para el desarrollo del SEP se tuvieron en cuenta dos aspectos fundamentales, la métrica a utilizar y la arquitectura del sistema.

## MÉTRICA

Para la estimación del grado de pertinencia se tuvieron en cuenta varios factores. Uno de ellos fue que el resultado final debía reflejar con claridad la pertinencia, es decir, que el valor les indicara realmente a los administradores y usuarios si la navegación fue pertinente o no. Para ello se tomaron como parámetros: el peso y el tamaño de la página que se extraen del log de DansGuardian.

DansGuardian para calcular el rating o peso de la página, asigna pesos positivos a las palabras no pertinentes y negativos a las pertinentes, se decidió de este mismo modo, clasificar como pertinentes aquellas páginas cuyo valor de rating fuese negativo y como no pertinentes las que tuvieran valor positivo, para no crear confusión con el criterio con que trabaja el DansGuardian. Mientras más alejado estén estos valores de cero más pertinente (tomando valores negativos) o menos pertinente (tomando valores positivos) será la página.



Figura 1: Escala de valores de pertinencia

Se decidió que el tamaño sólo tuviera influencia a la hora de evaluar páginas no pertinentes, ya que mientras mayor sean estas, mayor será el agravamiento en la calidad del servicio que se les brinda a los demás usuarios. Por lo tanto a la hora de calcular el grado de pertinencia, si la página es no pertinente y su tamaño está por encima de un valor máximo admitido, el resultado empeorará. La relación entre el peso y el tamaño se estableció por medio de los siguientes umbrales:

Tabla 1: Estimación de la pertinencia teniendo en cuenta el tamaño de las páginas no pertinentes.

Umbral	Tamaño de página no pertinente (Megabyte)		
	Menor que 1	Entre 1 y 10	Mayor que 10
Pertinencia de la página	rating	rating * 2	rating * 4

Para el caso de que la página sea pertinente, es decir que el rating obtenido del DansGuardian sea un valor negativo, no se aplicará ninguna modificación; la pertinencia de la página será igual al rating.

Un tercer parámetro que se tuvo en cuenta fue el horario en que se accede a las páginas no pertinentes, pues el acceso en horario laboral afecta a los demás usuarios que necesitan hacer un uso pertinente de la red. Este factor podrá ser o no utilizado para calcular la pertinencia de la página, se desarrolló para que funcionara de forma opcional, sólo si a la empresa le resultara útil. Se definió por defecto, para realizar este análisis, el horario de 8:00 am a 5:00 pm; pero también en dependencia de lo que defina la empresa podrá ser otro el horario. El método, teniendo en cuenta el horario, sustituye la Tabla 1 por la siguiente Tabla 2:

**Tabla 2: Estimación de la pertinencia teniendo en cuenta el tamaño y el horario en que se accedió a páginas no pertinentes.**

Umbral	Tamaño de página no pertinente (Megabyte)		
	Menor que 1	Entre 1 y 10	Mayor que 10
Pertinencia de la página fuera del horario crítico	Rating	rating * 2	rating * 4
Pertinencia de la página dentro del horario crítico	Rating	rating * 3	rating * 8

Luego de tener calculado el valor de pertinencia individual para cada una de las páginas visitadas, se necesitó un método para estimar el valor de pertinencia mensual del usuario. El siguiente ejemplo muestra el método utilizado para un grupo de tres usuarios.

**Tabla 3: Ejemplo de un conjunto de valores de pertinencia de páginas para un grupo de usuarios.**

Usuario	Valores de pertinencia de cada página										
X	-90	-80	-70	-60	-50	10	20	30	40	50	60
Y	-90	-80	-70	-60	-50	10	20	80	90	-	-
Z	-90	-80	-70	-60	-50	90	-	-	-	-	-

En la Tabla 3 se muestran las pertinencias de las páginas visitadas por tres usuarios. Para calcular la pertinencia total, además de hacer uso de los valores de pertinencia por página, se hace uso de la cantidad de páginas de cada tipo (pertinente o no pertinente), y se siguen los siguientes pasos:

1. Se calculan los promedios de páginas no pertinentes y pertinentes, Pnp y Pp, respectivamente.

Para X: Pp = -70, Pnp = 35 luego Pu(X) = -35

Para Y: Pp = -70, Pnp = 50 luego Pu(Y) = -20

Para Z: Pp = -70, Pnp = 90 luego Pu(Z) = 20

2. Se calculan las razones entre las cantidades de páginas pertinentes y no pertinentes, Cp y Cnp respectivamente. Siempre dividiendo la mayor cantidad entre la menor.

Para X: Cp = 5, Cnp = 6, luego Cnp / Cp = 1.2

Para Y:  $C_p = 5$ ,  $C_{np} = 4$ , luego  $C_p / C_{np} = 1.25$

Para Z:  $C_p = 5$ ,  $C_{np} = 1$ , luego  $C_p / C_{np} = 5$

3. Luego se toma, para cada usuario, el valor promedio que corresponde con la menor cantidad de páginas y se le divide la razón calculada anteriormente, obteniendo un valor modificado de promedio,  $P'$ . Esto tiene como objetivo disminuir la influencia del promedio correspondiente a la menor cantidad de páginas debido a que este pudiera modificar sustancialmente el resultado final evitando que se refleje la tendencia real del usuario.

Para X:  $P_p' = -70 / 1.2 = -58.33$

Para Y:  $P_{np}' = 50 / 1.25 = 40$

Para Z:  $P_{np}' = 90 / 5 = 18$

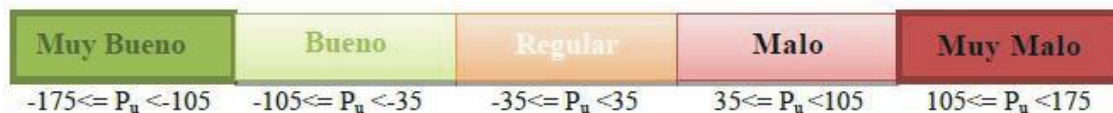
4. Por último se suman los promedios anteriores con los promedios que no fueron modificados, obteniendo la pertinencia total del usuario  $P_u$ . Para los usuarios de la Tabla 3, los resultados son los siguientes:

Para X:  $P_p' = -58.33$ ,  $P_{np} = 35$  luego  $P_u(X) = -23.33$

Para Y:  $P_p = -70$ ,  $P_{np}' = 40$  luego  $P_u(Y) = -30$

Para Z:  $P_p = -70$ ,  $P_{np}' = 18$  luego  $P_u(Z) = -52$

Por último, la métrica utilizada debe cumplir con un factor muy importante y es que los usuarios al ver su reporte de pertinencia entiendan que significa. La interpretación del grado de pertinencia mediante valores numéricos, como los obtenidos hasta el momento en el ejemplo, no brindan información de cuán pertinente o no fue el usuario. Para esto se utilizó un criterio que vincula los valores de pertinencia obtenidos ( $P_u$ ) con niveles de evaluación que sí emiten un juicio de comportamiento claro. Estos niveles son asignados del siguiente modo:



**Figura 2: Relación entre los valores de  $P_u$  y los criterios de pertinencia.**

Los valores del rango de la Figura 2 son hipotéticos, la determinación de los valores reales se explica más adelante. Luego, los usuarios "X" y "Y" le dieron un uso a Internet regular, y el usuario "Z" bueno, en cuanto a pertinencia se refiere. [2]

## ARQUITECTURA

El SEP es un complemento del Sistema Integrado de Contabilidad y Configuración (SICC), ya implantado en la CUJAE y en la Universidad de Cienfuegos. El lenguaje de programación que se escogió fue Python, debido a que es fácil de aprender y utilizar; además tiene un ciclo de desarrollo corto y facilita la comprensión del código por terceros. Otro motivo fue que es el lenguaje sobre el cual se desarrolló el SICC. En la Figura 3 se muestra como quedó estructurada la arquitectura del sistema.

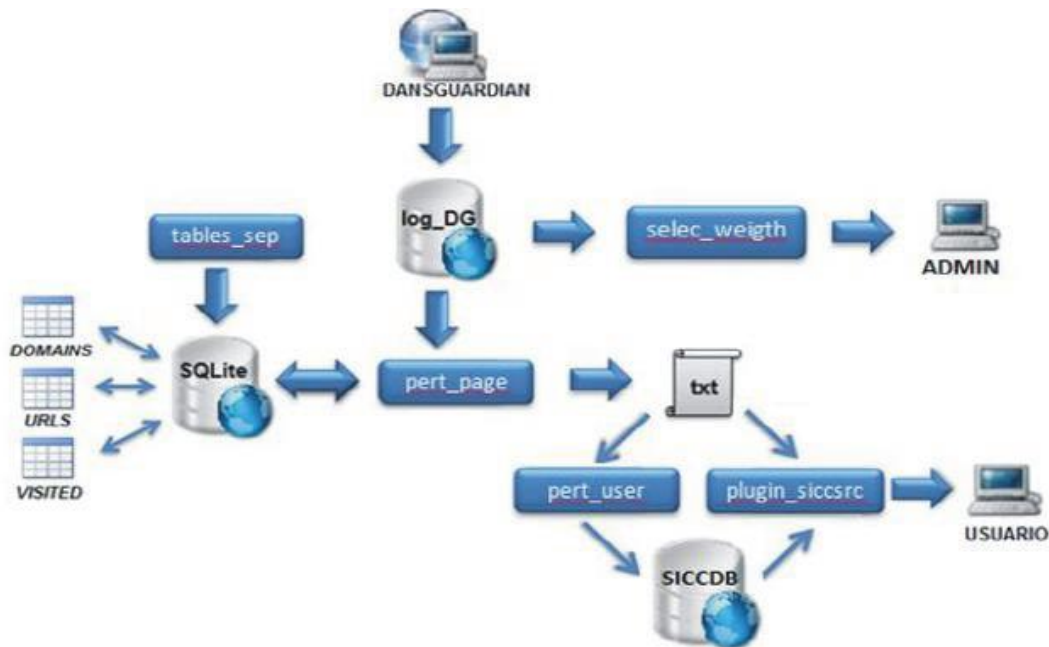


Figura 3: Arquitectura del SEP

En la base de datos SQLite se van a almacenar un conjunto de dominios y URLs para los cuales se va a tomar una decisión, en cuanto a peso se refiere, distinta a la que da DansGuardian. Este aspecto es muy importante porque el sistema necesita ser operable en cualquier entorno y muchas veces la valoración de las páginas que da DansGuardian no es la adecuada para determinada organización. La base de datos contiene tres tablas: DOMAINS, URLS y VISITED. En el caso que el administrador de una empresa desee hacer cambios en función a las necesidades y fines de su organización podrá hacerlo sin problemas a través de las tablas DOMAINS y URLS que contienen información de los sitios y sus pesos. Estas tablas podrán llenarse manualmente a través de la consola de SQLite. La tabla VISITED favorecerá en gran medida el desempeño del SEP a mediano y largo plazo. Puede darse el caso de que el rating calculado por DansGuardian sea cero debido a que al sumar los pesos de las palabras el resultado sea cero, o que DansGuardian no encuentre en la página ninguna de las palabras de sus listas. Al multiplicar este peso (cero) por el valor umbral que se determinó según el tamaño de la página el resultado es cero; entonces, el sistema no sabrá qué hacer con la página. Esto no es aceptable a la hora de valorar la pertinencia total porque el cero no aporta nada. Por tanto, en esta tabla se almacenarán las URLs de las páginas a las cuales DansGuardian de peso cero, así como la cantidad de accesos a las mismas. Esto le permitirá

conocer al administrador qué páginas son las más visitadas que no están clasificadas ni en DOMAINS, ni en URLS y que obtuvieron peso cero; para luego visitarlas y agregarlas en DOMAINS o en URLS con el peso adecuado para su empresa.

Inicialmente el administrador tendría que introducir en las tablas, los dominios y URLS a los cuales quiere que se les dé un tratamiento especial. Para facilitar el trabajo inicial se le incorporarán a las tablas DOMAINS y URLS todos los dominios y URLS del archivo bigblacklist en su versión más reciente. Esta bigblacklist es una gran lista ya definida que posee gran cantidad de información organizada por categorías como: sports, vacations, news, etc. La incorporación de los dominios y URLS a las tablas DOMAINS y URLS respectivamente, ahorra tiempo de clasificación manual. El proceso de llenado inicial de las tablas será muy sencillo gracias a la utilización del script tables\_sep. Este script tendrá un archivo de configuración en el cual estarán todas las categorías de la bigblacklist; bastará con introducirle peso a las categorías de interés y especificar la ruta de la bigblacklist, y luego al correr el script se llenarán las tablas.

Cuando se adquiera una nueva versión de la bigblacklist bastará con realizar el proceso anteriormente descrito y se actualizarán las entradas provenientes de la lista sin modificar las entradas agregadas manualmente por el administrador.

El script select\_weigth será utilizado cuando se inicie la implantación del SEP para recomendar al administrador los valores de peso que debe otorgar a las categorías de la bigblacklist. Este script toma del log que generó DansGuardian en la empresa todos los valores de rating; se recomienda que el log contenga información de al menos una semana para que la cantidad de muestras sea razonable. Tras un análisis matemático, el script devuelve un rango de valores de peso, parecido al de la Figura 4, dividido en cinco grupos: Muy Malo, Malo, Regular, Bueno y Muy Bueno. El administrador deberá ubicar cada una de las categorías de la bigblacklist en uno de estos grupos de acuerdo a la importancia para su organización, y asignarle un valor de peso que se corresponda con los límites que dividen al grupo.

Cuando ya están conformadas las tablas, el bloque pert\_page será el encargado de realizar el cálculo del grado de pertinencia por página y guardar los resultados. Este script interactúa directamente con el reporte o log del DansGuardian, con la base de datos SQLite y con un sistema de almacenamiento en archivos de texto. El sistema de almacenamiento en archivos de texto, consiste en una estructura de directorios de la forma \año\mes, en donde se almacenan todos los usuarios de forma tal que cada uno representa un archivo usuario.txt, que contiene las URLS visitadas con sus valores pertinencia calculados según la métrica especificada. Como parámetro inicial para el cálculo de pertinencia se toma el rating que otorga DansGuardian, siempre y cuando la URL o dominio accedido no esté registrado en las tablas DOMAINS o URLS, caso en el cual se utiliza el valor de rating que está asignado en estas tablas.

El módulo pert\_user es el encargado de calcular el valor de pertinencia mensual para cada uno de los usuarios. Esto lo hace leyendo de cada uno de los archivos de usuario, los valores de pertinencia de cada página visitada y aplicando el método para el cálculo de pertinencia mensual, obtiene los valores totales para cada usuario y los almacena en la SICCDB4. [2]

Para mostrar los reportes por usuarios se utilizó la herramienta Subsistema de Reportes de Consumo del SICC-IP (SICCSRC). El SICCSRC es un sistema para obtener y mostrar información distribuida de consumo de los servicios ofrecidos en una red IP. Su diseño permite abstraer al desarrollador de un nuevo servicio, de los detalles de comunicaciones y presentación de los datos al usuario [8]. El SICCSRC trae,

junto con sus librerías, un plug-in para el reporte del consumo del servicio de Internet que consiste en dos scripts, `internet_resume` para mostrar la información de resumen que se extrae de la SICCDB, e `internet_details` para mostrar los detalles del consumo. Para el desarrollo del SEP se agregó a estos scripts la información de resumen y de detalles de pertinencia, de modo que el plug-in para el servicio de Internet también mostrará esta información.



## CONCLUSIONES

No existen en la actualidad herramientas que evalúen el tráfico de Internet en más o menos pertinente y la necesidad de algún tipo de software que realice esta función es imperante. Los sistemas de reporte y los filtros de contenido están concebidos con otros objetivos que no están vinculados de ninguna manera a la evaluación de pertinencia. Aunque es cierto que se pudiera realizar algún estimado de pertinencia mediante el análisis manual de las bitácoras que generan este tipo de herramientas, esto no sería factible, ni recomendable, por el gran trabajo y dedicación que requeriría de los administradores, debido a lo complejos y extensos que pudieran llegar a ser estos reportes. Sin embargo haciendo uso de la técnica de clasificación por ponderación de palabras utilizadas por los filtros de contenido como DansGuardian, sí se pudiera estimar el grado de pertinencia de una red empresarial.

La métrica fue escogida de forma tal que brindara información real del grado de pertinencia, teniendo en cuenta no sólo los valores de rating, sino también otros factores como: si la cantidad de páginas que predominan son o no pertinentes, el tamaño de la páginas y el horario en el cual se accede a Internet. El SEP fue diseñado con la intención de que pudiera ser implantado en cualquier empresa, brindando la facilidad de modificar sus componentes de forma tal que respondiese a las necesidades y objetivos de esta. Se ofreció la opción de usar información categorizada en la bigblacklist, para lograr un mejor rendimiento y un ahorro considerable de tiempo de clasificación manual. Además, mediante la integración de una tabla de sitios no clasificados (VISITED) se logró que el sistema pudiera ser mejorado considerablemente con el tiempo.

## Referencias Bibliográficas

1. "Control interno". <http://fjavierrama.wordpress.com/auditoria-informatica/control-interno/>
2. PÉREZ MANTILLA, Susana; GUTIÉRREZ LÓPEZ, Aryam. "Herramienta para el análisis de la pertinencia de la navegación por internet de una red empresarial". Tesis para optar por el título de Ingeniero en Telecomunicaciones y Electrónica, CUJAE, Ciudad de La Habana, 2012.
3. "SARG ". <http://www.ecured.cu/index.php/SARG>. 2012
4. "Internet Access Monitor (Libro blanco técnico)".  
[http://www.redline-software.com/spa/products/iam/white\\_papers/IAM.pdf](http://www.redline-software.com/spa/products/iam/white_papers/IAM.pdf)
5. "SurfCop (Libro blanco técnico)".  
[http://www.redline-software.com/spa/products/surfcop/white\\_papers/SurfCop.pdf](http://www.redline-software.com/spa/products/surfcop/white_papers/SurfCop.pdf)
6. PUERTAS, Enrique; GÓMEZ HIDALGO, José María; CARRERO, Francisco; DE BUENAGA, Manuel. "Filtrado de contenidos web en español dentro del proyecto POESIA"
7. MARTIN MALDONADO, Daniel. "Analizando el contenido de las páginas web con DansGuardian".  
<http://www.aplicacionesempresariales.com/analizando-el-contenido-de-las-paginas-web-con-dansguardian.html>. 2008
8. MACHADO PIÑEIRO, Diana. "Manual de instalación del Subsistema de Reportes de Consumo del SICC-IP (SICCSRC)". 2008