Universidad de las Ciencias Informáticas

Facultad 10

Propuesta para la prevención y detección de sitios anonimizadores

Trabajo de Diploma para optar por el Título de Ingeniero en Ciencias Informáticas

Autor: Daileny Hernández Barreiro

Tutor: Ing. Luis Enrique Sánchez Arce

Co-Tutor: Msc. Damaris Cruz Amarán

Ciudad de La Habana, 15 de junio de 2010

Declaración de autoría

Declaramos ser los únicos autores de la presente tesis y reconocemos los derechos patrimoniales de la misma a la Universidad de las Ciencias Informáticas (UCI), con caracter exclusivo.

Para que así conste firmamos la presente a los ____ días del mes de _____ del año _____.

_____

Daileny Hernández Barreiro

Autor

_____ _____

Ing. Luis Enrique Sánchez Arce — Msc. Damaris Cruz Amarán

Tutor — Co-Tutor

Agradecimientos

A mi mamá y mi hermana por ser mis guías, por ser mi ejemplo a seguir, por apoyarme en todo momento y por ser la fuerza que me impulsa cada día a tratar de superarme y crecer como persona.

A mis abuelos Nena y Carlos por creer tanto en mí y por todo el amor que me han dado.

A Amambay por ser la persona que incondicionalmente ha estado a mi lado estos 5 años, por todo el apoyo y el aliento que me ha dado, por ser tan buen amigo, a él le debo todo lo que soy.

A Lili y Lisi por ser las mejores amigas del mundo y brindarme toda su confianza y cariño.

A Reglita por todos sus consejos y cuidarme a mi mamá y mi hermana cuando yo no estaba.

A todos mis amigos del aula y del polo productivo SINI por todos los momentos lindos que pasamos, por enseñarme tantas cosas y brindarme su ayuda cada vez que la necesité.

A todos mis amigos de la universidad que aunque no los mencione siempre estarán en mi corazón.

A todos mis profesores por su ayuda en mi preparación profesional.

A mi tutor Luis por sus orientaciones y la confianza que depositó en mi.

A la Revolución Cubana por darme la oportunidad de estudiar en esta universidad tan linda.

Daileny.

Dedicatoria

Dedico este trabajo de diploma a mi mamá Mari Isel y mi hermana Daily, ustedes son mis estrellas, este logro también es de ustedes por ser mi fuente de inspiración, las quiero mucho.

Daileny.

“Saber no es suficiente; tenemos que aplicarlo. Tener voluntad no es suficiente: tenemos que implementarla”

Johann Wolfgang von Goethe

Resumen

En la UCI se desarrolla, desde el año 2005 y a petición de la Oficina de Seguridad para las Redes Informáticas, un sistema de Filtrado de Paquetes por Contenido que pretende regular (aceptar o denegar) el acceso de usuarios a determinados contenidos de Internet y brindarles una navegación segura, adaptándose a las normas y políticas de las instituciones en que pueda instalarse. El uso de sitios anonimizadores en muchas ocasiones permite evadir las políticas de seguridad establecidas por este producto en la navegación de los usuarios por Internet, provocando que constituyan una eminente amenaza que atenta contra el correcto funcionamiento de este sistema. Una de las vías más utilizadas para la detección del uso de estos proxies es mediante las listas preclasificadas que brindan numerosos sitios, las cuales recopilan sitios anonimizadores ya conocidos y que se encuentren en funcionamiento, pero el constante crecimiento de Internet impide que esta técnica sea eficiente, pues estos sitios cambian constantemente y surgen otros nuevos, lo cual imposibilita que dichas listas estén completamente actualizadas, dando paso entonces a la necesidad de conocer e investigar acerca de otras maneras para detectar automáticamente (en tiempo real), el uso de estas herramientas que posibilitan la navegación anónima.

En este trabajo se realiza un estudio de las técnicas existentes para este fin y se proponen otras resultado de la presente investigación. También se implementa un prototipo funcional para la detección de sitios anonimizadores, al cual se le realizaron pruebas que mostraron un correcto funcionamiento del mismo, demostrando así que puede ser satisfactoriamente integrado al producto Filtrado de Paquetes por Contenidos (Filpacon).

Índice general

Índice de figuras

Introducción

Internet, la infraestructura de redes más grande a escala mundial, es actualmente uno de los pilares de la de información con mayor impacto para la sociedad. Si bien es una fuente de conocimiento de incuestionable valor, al albergar contenidos educativos, culturales, científicos, informativos, recreativos, también puede resultar sumamente dañina al contener otros materiales que pueden resultar inadecuados, ilícitos, o incluso ilegales para algunos países, como pueden ser la pornografía, el terrorismo, el racismo, la pedofilia, entre otros.

El hecho de que los usuarios accedan a este tipo de información puede traer grandes consecuencias para las empresas y/u organizaciones. El aprovechamiento del horario laboral de los empleados se ve grandemente afectado, disminuyendo la productividad de los mismos y aumentando los costos. Se expone la red a programas maliciosos, virus, malware y otras amenazas que atentan contra la seguridad de la organización, inclusive, se puede ver afectada la reputación de la empresa o la organización, que podría costar años reparar.

Desde que se detectó este serio problema, controlar el acceso a estos contenidos se ha convertido en una problemática. Muchos han sido los intentos que se han hecho por parte de los gobiernos para este fin, como la creación de diversas leyes que penalizan la publicación y el acceso a estos tipos de información, pero debido al anonimato que ofrece la gran red de redes para publicar y acceder a los mismos, han provocado que estas legislaciones no tengan mucho éxito.

El uso de los Sistemas de Filtrado de Contenido es, hasta el momento, la mejor solución que se ha encontrado o aplicado. Estos sistemas regulan el acceso a información que se considere inadecuada, ilícita o nociva, en dependencia de ciertas reglas y parámetros que son configurados previamente por los administradores del sistema.

La navegación anónima, específicamente algunos recursos que la hacen posible, como lo son los sitios anonimizadores o sitios proxies, se ha convertido en una de las vías que atenta contra la efectividad de los sistemas de filtrado, ya que en muchas ocasiones permite evadir las políticas de seguridad establecidas en la navegación de los usuarios por Internet, provocando así que constituya una eminente amenaza contra el correcto funcionamiento de este tipo de software que propicia la regulación del acceso a los contenidos en Internet.

Para la detección del uso de los sitios anonimizadores comúnmente se utilizan listas preclasificadas, las cuales recopilan sitios proxies ya conocidos y que se encuentren en funcionamiento, pero el constante crecimiento de Internet impide que esta técnica sea eficiente, pues estos sitios web cambian constantemente y surgen otros nuevos, lo cual imposibilita que dichas listas estén completamente actualizadas y que cuenten con todos los sitios anonimizadores existentes, dando paso entonces a la necesidad de conocer e investigar acerca de otras maneras para detectar automáticamente (en tiempo real), el uso de estas aplicaciones.

Actualmente en la UCI se está desarrollando un sistema de filtrado, llamado Filpacon. Este es un nuevo producto que cuenta con varias características que lo hacen sobresalir en el mercado frente a soluciones de su mismo tipo. A pesar de todo esto Filpacon no queda exento de la amenaza que representan los sitios anonimizadores, puesto que no tiene incorporada ninguna herramienta que detecte el uso por parte de los usuarios de dichos sitios que permiten la navegación anónima.

Centrándose en la situación que presenta el producto antes mencionado, se define el **problema científico** de este trabajo: ¿Cómo detectar y prevenir de forma automática el uso de sitios anonimizadores durante el proceso de filtrado del producto Filpacon?

Se puede formular entonces la siguiente **idea a defender:** Mediante la investigación de técnicas para detectar el acceso a sitios anonimizadores y el estudio de estos sistemas, es posible la concepción de una solución que impida el uso por parte de los usuarios de dicho tipo de sitio web, permitiendo evitar lo que pudiera constituir una brecha de seguridad en el producto Filpacon

La estrategia para resolver el dilema planteado está rectorada por el **objetivo general** el cual se precisa como: Definir los mecanismos para la detección y prevención de sitios anonimizadores. De aquí se obtienen varios **objetivos específicos**:

- Identificar las principales vías de detección de sitios anonimizadores conocidos y no conocidos.
- Identificar otros sistemas que pueden ser utilizados como sitios anonimizadores, no siendo este su fin específico.
- Desarrollar un prototipo funcional para detectar de forma automática el uso de sitios anonimizadores.

El cumplimiento y ejecución de los objetivos propuestos lleva consigo un estudio que sirva de apoyo para el cumplimiento de lo antes mencionado, es por ello que se define como **objeto de estudio** *el uso de los sitios anonimizadores*; quedando como **campo de acción** *el proceso de detección de los sitios anonimizadores.*

Para dar cumplimiento a los objetivos planteados se han definido las siguientes tareas de investigación:

- Realizar un estudio y análisis de los métodos conocidos acerca de la detección de sitios anonimizadores.
- Realizar análisis de las herramientas existentes para la detección del uso de sitios anonimizadores.
- Definir las tecnologías a emplear para concebir un prototipo funcional que se integre con el sistema Filpacon.
- Desarrollar un prototipo funcional para detectar de forma automática el uso de sitios anonimizadores

En el cumplimiento de las tareas se usarán los siguientes métodos teóricos:

- El **Analítico-Sintético** se aplicará para entender los sitios proxies a partir del análisis de las características que presentan y observar la evolución de los principales sistemas utilizados en la construcción de este tipo de sitio web. Además servirá de base para formular conclusiones a través de la síntesis de los conocimientos y resultados obtenidos.
- El **Histórico-Lógico** permitirá que se analice el desarrollo histórico del objeto de estudio y se encuentre la lógica interna del desarrollo.
- La **Modelación** se utilizará para diseñar un prototipo funcional que refleje lo mejor posible la realidad que presenta el proceso de detección de un sitio anonimizador.

La presente investigación se encuentra estructurada en el documento en dos capítulos, a continuación se brinda un breve resumen de los aspectos que se abordaron en cada uno.

- **Capítulo I: Navegar de incógnito en la web: Sitios Anonimizadores.** En este capítulo se realiza un estudio de algunas soluciones existentes con respecto a la detección de sitios anonimizadores. Se abordan temas relacionados con los sitios anonimizadores que ayudan a ampliar los conocimientos acerca de este recurso, se describen técnicas utilizadas para la detección de este tipo de intermediario de la red así como servicios que pueden ser usados como sitios anonimizadores sin ser este su fin, como es el caso de los traductores de idiomas.
- **Capítulo II: Propuesta de Solución.** En este capítulo se describe otra técnica resultado de la presente investigación para detectar el uso de sitios anonimizadores, la cual se basa en la búsqueda de determinados patrones en el código HyperText Markup Language (HTML) de la página. Se describe la integración que tendrá con el sistema Filpacon la solución propuesta, presentando además la nueva arquitectura que adquiriría este software. Se proponen todos los aspectos a tener en cuenta para la implementación de un prototipo funcional y, por último, se comprueba mediante pruebas, la efectividad con que cuenta el mismo después de su implementación.

Capítulo 1

Navegar de incógnito en la web: Sitios Anonimizadores

Los sitios anonimizadores permiten y posibilitan la navegación anónima, haciendo simple a los usuarios violar las políticas de acceso y seguridad impuestas por los Sistemas de Filtrado de Contenido para la exploración segura de la web. En el marco de los Filtros de Contenido contar con herramientas que permitan la detección del uso de este tipo de intermediario de la red se ha vuelto indispensable.

Tener dominio sobre estas aplicaciones, sus características y funcionamiento, son premisas imprescindibles que ayudarán a sentar las bases teóricas para llegar a originar una solución como la se requiere.

1.1. Trabajos Similares

Ante la gran necesidad de que el sistema Filpacon cuente con una herramienta que le posibilite detectar el uso de sitios anonimizadores en tiempo real, se hace necesario el estudio de otras soluciones existentes que permitan detectar este tipo de acceso por parte de los usuarios que utilicen el sistema.

1.1.1. Ámbito Internacional

La mayoría de los Sistemas de Filtrado de Contenido cuentan con una robusta defensa ante el uso de los sitios anonimizadores. Un sistema de gran prestigio en este ámbito es **eSafe**, este producto además de filtrar el contenido web, ofrece protección anti-virus capaz de detectar troyanos, gusanos o cualquier otro

tipo de virus e incluye también un *firewall* que controla el acceso a Internet y la existencia de software espía [1]. Cuenta con una herramienta llamada **Anti-Anonymizer** la cual es considerada como una de las soluciones más completas para combatir la tecnología de los anonimizadores, de manera que protege la inversión de seguridad y limita la responsabilidad del cliente debido al mal uso de los recursos que puedan realizar sus empleados. eSafe Anti- Anonymizer utiliza tres capas de protección para evitar la utilización de anonimizadores:

- **eSafe Web:** Solución de filtrado de Uniform Resource Locator (URL) que bloquea el acceso a anonimizadores conocidos y sitios web no reconocidos.
- **eSafe Applifilter:** Obliga a utilizar un proxy autorizado mientras bloquea los intentos de redireccionamiento de puertos para acceso privado.
- **eSafe Web SSL:** Aplica la política sobre sitios anonimizadores cifrados Secure Socket Layer (SSL). Puede identificar y bloquear de forma proactiva los anonimizadores desconocidos así como bloquear sitios SSL con firma propia ("anonimizadores caseros").

eSafe Web SSL tiene la capacidad de validar certificados, políticas, autores, revocaciones entre otros. Mediante la utilización de **eSafe Anti-Anonymizer**, los especialistas del equipo Content Security ResponseTeam (CSRT) de Aladdin[1] bloquearon el 100 % de los anonimizadores en distintas pruebas de laboratorios y comprobaron que mantiene los equipos libres de virus, software espía, gusanos y otros ataques malintencionados. Las soluciones de la competencia no resistieron los ataques proxy y permitieron que los usuarios abandonaran la red protegida.

iPrism Web Filter es otra alternativa para el filtrado de la web. Protege a las organizaciones de las amenazas provenientes del Internet con una tecnología robusta para el filtrado dirigido por políticas y reportes precisos. Sus principales características son [2]:

- **Protección Antivirus/Malware:** Bloqueo de *spyware*, *malware*, peer-to-peer (P2P) en el perímetro, e incluye un poderoso motor antivirus de cuatro factores.

[1] http://www.aladdin.com/

- **Autentificación:** Autentificación vía el Directorio Activo y Lightweight Directory Access Protocol (LDAP).
- **Filtrado Móvil y Remoto:** Extiende fácilmente la ejecución de políticas de seguridad a usuarios móviles o con laptops.
- **Administración Delegada:** Provee la asignación de tareas de administración dentro de la organización.
- **Fácil de Implementar:** Es fácil de instalar en una amplia variedad de ambientes de redes.
- **Administración Flexible de Políticas:** Definición de políticas flexibles globales, por grupo, usuario o Dirección IP (IP) con más de 80 categorías de donde seleccionar.
- **Reportes Integrales con Despliegue de Cambio Rápido:** Reportes detallados en tiempos reales disponibles en la caja.
- **Defensa para Anonimizadores:** Defensa robusta para anonimizadores con bloqueo dinámico de proxies.

Este sistema de filtrado protege las organizaciones de los anonimizadores y del daño que pueden causar con un acercamiento multicapa que lo protege en el perímetro. Una línea de defensa es la poderosa base de datos iGuard [2] que es revisada al 100 % por humanos, la cual incluye miles de sitios anonimizadores que se pueden bloquear con un simple clic. Y debido a que la base de datos iGuard se actualiza constantemente, los nuevos sitios anonimizadores son identificados y descargados todos los días.

Además de la revisión humana, este sistema agrega otra capa de protección en contra de los anonimizadores. Emplea agentes con inteligencia artificial en su sistema operativo para analizar los patrones en las peticiones de URLs. Si durante el análisis de datos se detecta un patrón sospechoso, dinámicamente bloquea el acceso a ese sitio. El equipo de iGuard también identifica activamente los patrones únicos y consistentes para asistir en la clasificación dinámica en tiempo real. La lista de patrones actuales del iGuard ofrece una excelente cobertura de paquetes de proxy tal como el PHPProxy y el CGIProxy [3]. La actualización a estos

[2] http://www.iguard.org//

[3] Scripts generadores de sitios proxies

patrones es publicada y distribuida cada hora en las actualizaciones críticas para todos los clientes de **iPrism** [3].

Sophos Web Security and Control es otro software que se encarga del control del acceso a Internet para una navegación segura y productiva, además de ofrecer servicios de antivirus bloqueando programas espías, virus, robo de información, programas maliciosos y aplicaciones no deseadas en la puerta de enlace [4]. Como parte de sus principales cualidades se encuentra que posee detección automática de proxies anónimos. Para ser eficaz en este aspecto, su solución esta compuesta por dos elementos: la detección de proxies anónimos mediante listas existentes en Internet que recogen esta información y la detección en tiempo real de dichos sitios [5].

- **Detección de proxies anónimos mediante listas existentes en Internet:** Monitorea foros de Internet para localizar sitios anonimizadores, publicando las actualizaciones de la lista de bloqueo cada 15 minutos.
- **La detección en tiempo real:** Automáticamente inspecciona el tráfico en busca de pistas de enrutamiento a través de un proxy anónimo.

Mediante estos dos componentes proporciona la protección adecuada de la organización y el filtrado de todo el tráfico sobre la base de criterios establecidos.

La detección de sitios anonimizadores, importante cualidad que poseen los sistemas mencionados, no son alternativas que puedan adaptarse a Filpacon. Esto se debe a que son soluciones privativas, y por tanto no pueden reutilizarse para ser moldeadas y aprovechadas.

1.1.2. Ámbito Nacional

En el ámbito nacional Filpacon en la única solución concebida de su tipo, imponiendo con esto gran novedad entre sus componentes y apareciendo con esto la necesidad de buscar alternativas que se encarguen de hacerlo un software de excelencia, contando con todas las características que lo hagan ser parte de los productos de punta en el marco de los Sistemas de Filtrado de Contenido.

Debido a que no se cuenta con ninguna solución que se pueda aprovechar para suplir la necesidad de detectar el uso de los sitios anonimizadores, se hace ineludible desarrollar una herramienta que supla esta carencia.

1.2. Navegación Anónima en Internet, Anonimizadores

Hay muchos motivos por los cuales se decide navegar de manera anónima por Internet. Uno de ellos es ocultar la dirección IP. Cuando se visita un sitio web la máquina que alberga dicho sitio puede obtener gran cantidad de información relacionada con la Computadora Personal (PC) desde la cual se accede a el, como por ejemplo su dirección IP.

Teniendo esta dirección es posible determinar no solo el país de procedencia, sino también la ubicación más precisa del origen de la conexión. Aunque no es fácil conectar la dirección IP con determinada persona, a veces simplemente no es deseado por los usuarios para exponer su posición o las propiedades de su navegador. A la par de esto se fueron sumando otras causas para tomar esta conducta en el buceo de la red de redes. Cada movimiento que realiza el usuario queda grabado por sus proveedores de Internet y por los servidores que hospedan los sitios web que visitan. Esta información es almacenada y a menudo entregada a terceras partes, las cuales la usarán para su propio beneficio (por ejemplo mostrar avisos personalizados, comenzar campañas de marketing dirigidas).

Otra razón por la cual este tipo de navegación se ha hecho tan popular es que puede evitar los filtros de la puerta de enlace, trayendo consigo que los contenidos y las infecciones maliciosas que a veces transportan puedan penetrar la red, siendo esta una manera de vencer las medidas de seguridad y permitiendo acceder a los internautas a sitios bloqueados, quedando fuera con esto de cualquier restricción sobre los contenidos a los cuales ambicionaran acceder. Debido a todo esto, navegar anónimamente en la web se ha ido convirtiendo en una tendencia en los últimos años, para hacerla un hecho se utilizan diversas herramientas y servicios que se especializan en este sentido.

Uno de los servicios líderes en el mundo en cuanto la navegación anónima es la red The Onion Ring (TOR). Es un sistema que no sólo permite la navegación anónima, sino también P2P, e-mail y chat Internet Relay

Chat (IRC) anónimos. EL objetivo como se describe en la página web del proyecto TOR es: "... transmitiendo sus comunicaciones en torno a una red distribuida de repetidores llevados por voluntarios de todo el mundo: evita que alguien que observa su conexión a Internet aprenda qué sitios visita, y evita que descubran su ubicación física mediante los sitios que visita."

Esta herramienta protege a los usuarios trasmitiendo sus comunicaciones a través de una red distribuida de repetidores llevados por voluntarios de todo el mundo. Funciona en muchas de las aplicaciones existentes, incluyendo los navegadores web, clientes de mensajería instantánea, acceso remoto, y otras aplicaciones basadas en el protocolo Transmission Control Protocol (TCP). Esta red es usada por cientos de miles de personas en todo el mundo [6]. Conjuntamente son utilizados diversos navegadores anónimos como es el caso de OperaTor[4] y XeroBank Browser[5].

Otra forma de navegar anónimamente consiste en la utilización de servidores intermediarios que accedan a los sitios web y retribuyan las páginas que se deseen visitar. Estos servidores son denominados proxies.

1.3. Servidor Proxy

Un proxy en términos de informática es un sistema intermediario entre *hosts*[6] internos de una red de área local y los *hosts* de Internet de forma tal que reciba las requisiciones de unos y se las pase a los otros previa verificación de accesos y privilegios [7]. Estos sistemas pueden correr en *hosts* "dual-homed", es decir, máquinas conectadas a dos redes pero que no tienen capacidad de enrutamiento, o en *hosts* "bastión", los cuales tienen como función ser el punto de contacto de los usuarios de la red interna con otro tipo de redes, siendo los encargados de filtrar el tráfico de entrada y salida y de ocultar la configuración de la red hacia afuera.

La comunicación entre el programa cliente y el servidor proxy puede realizarse de dos formas distintas:

[4] http://archetwist.com/en/opera/operator

[5] https://xerobank.com/download/xb-browser/

[6] El término host es usado en informática para referirse a los computadores conectados a la red, que proveen o utilizan servicios a/de ella.

- **Custom Client Software:** El cliente debe saber como opera el servidor proxy, como contactarlo, como pasar la información al servidor real. Se trata de un software cliente *standard* que ha sido modificado para que cumpla ciertos requerimientos.
- **Custom User Procedures:** El usuario utiliza un cliente *standard* para conectarse con un servidor proxy y usa diferentes procedimientos (comandos del servidor proxy) para pasar información acerca del servidor real al cual quiere conectarse. El servidor proxy realiza la conexión con el servidor real.

Los proxies se agrupan en dos grandes clasificaciones: los servidores proxies de SOCKS[7] y los servidores proxies de aplicación. Los servidores proxies de SOCKS se parecen bastante a un panel de conmutación. Tan sólo establecen la conexión entre su sistema y otro sistema externo. La mayoría de los servidores SOCKS presentan el inconveniente de que sólo trabajan con conexiones del tipo TCP y como cortafuegos no suministran autenticación para los usuarios. Sin embargo, su ventaja es que registran los sitios a los que cada usuario se ha conectado.

Por otro lado, los proxies de aplicaciones son servidores que conocen sobre una aplicación en particular y proveen servicios proxies para ella. Entienden e interpretan comandos de un protocolo en particular. Con este tipo de servidores es necesario contar con uno de ellos para cada servicio. Reciben también el nombre de servidores dedicados.

Un ejemplo de este tipo de proxy es el proxy Hypertext Transfer Protocol (HTTP) o proxy web como más comúnmente es conocido. Su tarea principal es interceptar la navegación de los clientes por páginas web, respondiendo con esto a varios motivos ya sean por seguridad, rendimiento o incluso por anonimato.

Funcionamiento de un servidor proxy

El principio de funcionamiento de los proxies comienza cuando un programa cliente del usuario se comunica con ellos enviando un pedido de conexión con un servidor real. El servidor proxy evalúa esta requisición y decide si se permitirá la conexión. Si el servidor proxy permite la conexión, envía al servidor real la solicitud

[7] SOCKS es un protocolo de Internet que permite a las aplicaciones cliente-servidor usar de manera transparente los servicios de un firewall de red.

Figura 1.1: Funcionamiento de un servidor proxy

recibida desde el cliente. De este modo, un servidor proxy se ve como un servidor cuando acepta pedidos de clientes y como cliente cuando envía solicitudes a un servidor real. Una vez establecida la comunicación entre un cliente y un servidor real, el servidor proxy actúa como un retransmisor pasando comandos y respuestas de un lado a otro. Un punto importante a tener en cuenta en este tipo de conexión es que es totalmente transparente. Un usuario nunca se entera de que existe un intermediario en la conexión que ha establecido. En la Figura 1.1 uno se ilustra el funcionamiento de este tipo de servidores.

Cuando el proxy recibe el mensaje de petición del cliente puede generar una respuesta propia o retransmitirlo al servidor final. En el segundo caso, el proxy puede introducir modificaciones en la petición, según la aplicación para la que esté diseñado y cuando reciba la respuesta del servidor final, la retransmitirá al cliente, también con la posibilidad de efectuar cambios. Asimismo, es posible que haya más de un proxy en la cadena, es decir, que el primer proxy no se conecte directamente al servidor final, sino por medio de otro proxy y así sucesivamente.

En el caso de los proxies web, en las peticiones que un cliente (u otro proxy) envía a un proxy, existe una variación respecto al caso de la conexión directa de cliente a servidor: el Uniform Resource Identifier (URI) de la línea de petición no debe ser un URL HTTP relativo, sino debe ser un URI absoluto. De lo contrario, el proxy no sabría cuál es el servidor final que va destinada la petición.

El uso de un servidor proxy tiene diferentes aplicaciones. Las principales son las siguientes:

- **Actuar como cortafuegos que aísle una red local con el resto de las redes:** En esta configuración los clientes no tienen acceso directo al exterior de su red, y toda comunicación con los servidores remotos tiene lugar por medio del proxy. Ello permite minimizar el riesgo de que usuarios externos comprometan la seguridad de los sistemas locales mediante accesos no autorizados, sabotajes, entre otros.
- **Tener una memoria caché compartida entre los usuarios de la red local:** Si diferentes clientes solicitan directamente el mismo recurso, por norma general guardarán la misma copia de la respuesta en sus respectivas memorias caché. Si lo solicitan por medio de un proxy, la primera petición necesitará acceso al servidor remoto, sin embargo, las siguientes quizás puedan aprovechar la copia ya guardada en la caché.
- **Construir una jerarquía de memorias caché de proxies:** en el nivel más bajo se encuentran los proxies a que acceden directamente los clientes, en un segundo nivel existen los proxies a que acceden los del primer nivel y así consecutivamente. Incluso puede haber proxies a escala de todo un país. Existe un protocolo denominado Internet Cache Protocol (ICP) que permite coordinar los diferentes servidores proxy de una jerarquía.

Ventajas que ofrecen los servidores proxy

Entre las ventajas que se derivan del uso de este tipo de herramientas se encuentran:

- **Autenticación**: Debido a que el proxy es una herramienta intermediaria indispensable para los usuarios de una red interna que quieren acceder a recursos externos, se utiliza para autenticar usuarios, es decir, pedirles que se identifiquen con un nombre de usuario y una contraseña para navegar en Internet.
- **Control de contenidos**: Al utilizar un servidor proxy, las conexiones de la red interna pueden rastrearse al crear registros de actividad (*logs*) para guardar sistemáticamente las peticiones de los usuarios cuando solicitan conexiones a Internet. Gracias a esto, las conexiones de Internet pueden filtrarse al analizar tanto las solicitudes del cliente como las respuestas del servidor. El filtrado que se realiza comparando

la solicitud del cliente con una lista de solicitudes autorizadas se denomina lista blanca; y el filtrado que se realiza con una lista de sitios prohibidos se denomina lista negra. Finalmente, el análisis de las respuestas del servidor que cumplen con una lista de criterios (como palabras clave) se denomina filtrado de contenido. Más concretamente el funcionamiento consiste en denegar el acceso a nombres de dominio o direcciones Web que contengan patrones en común. Ejemplos: pornografía, violencia, terrorismo. Así por ejemplo, si un usuario intenta ingresar a una página cuyo contenido esté relacionado con los patrones antes mencionados, se denegará el acceso a dicha página.

- **Control de ancho de banda de las descargas**: Con el control de ancho de banda se puede manejar de una manera más eficiente la tasa de transferencia de cada uno de los usuarios, de esta manera, se puede evitar que un usuario use todo el ancho de banda dejando a los demás usuarios con una tasa de transferencia lenta y deficiente.

- **Reportes estadísticos de la navegación**: Esta herramienta permite saber dónde han estado navegando los usuarios en Internet, pudiendo saber qué usuarios accedieron a qué sitios, a qué horas, cuantos *bytes* han sido descargados, relación de sitios denegados, errores de autentificación entre otros. Esta función es muy importante, principalmente para las empresas que quieren tener un control de accesos y ancho de banda de acceso a Internet.

- **Restricción del tiempo de navegación**: Por medio del proxy se puede tener el control de tiempo de navegación, restringiendo el horario en el cual determinados usuarios en la red interna pueden navegar en Internet.

Además de los proxies web, existe una gran variedad de proxies en dependencia de la finalidad que se persigue como por ejemplo los proxies Nat, los proxies transparentes, los proxies inversos, entre otros. Para navegar de incógnito en Internet, los proxies anónimos son los más utilizados.

1.4. Servidor Proxy Anónimo

Un proxy anónimo es un software que se usa para reenviar peticiones web después de ocultar la dirección de origen para los usuarios que quieren navegar en forma anónima [7]. En la Figura 1.2 se ilustra más

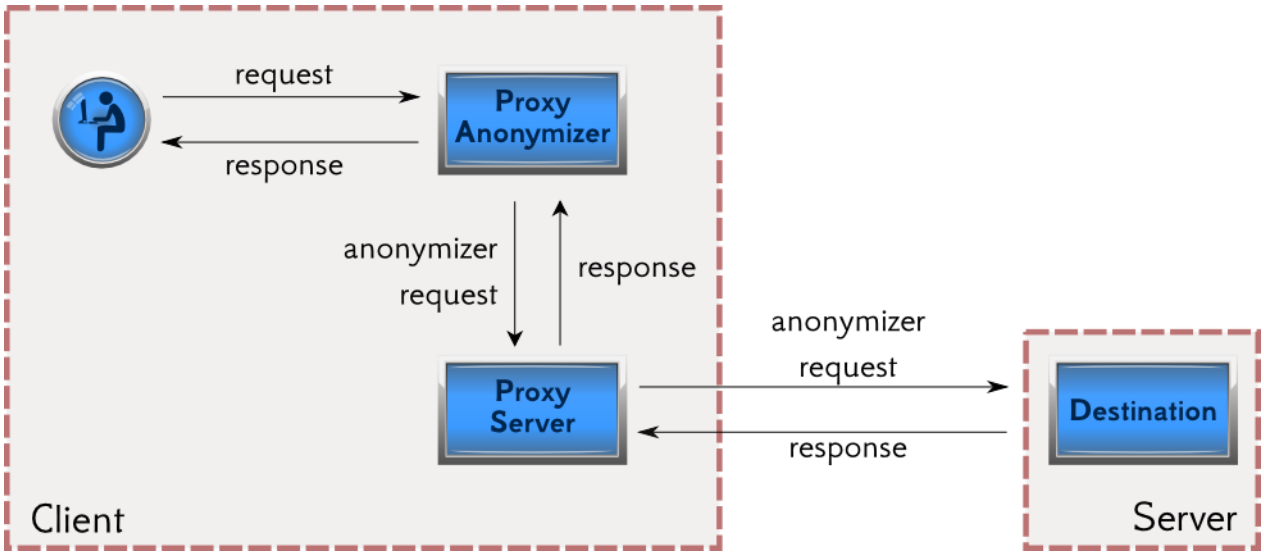

Figura 1.2: Funcionamiento de un proxy anónimo

claramente su funcionamiento. Su denominación de anónimos se debe a que permiten navegar al usuario bajo su IP pública. Este tipo de proxy es también muy utilizado en la seguridad informática, en las actividades de hackeo de sitios web, servidores, así como en el ingreso de bases de datos *online*. Una variante de los proxies anónimos son los sitios web anonimizadores, siendo estos el eje principal de este trabajo.

1.4.1. ¿Qué son los sitios web anonimizadores?

Son sitios web que brindan el mismo servicio que los proxies anónimos, comúnmente se les conoce como sitios proxies. Para crear el efecto de navegar anónimamente estos sitios son los preferidos por la mayoría de los usuarios, debiéndose esto a que no requieren de ninguna configuración previa del navegador ni de la instalación de ningún software adicional. Actúan como proxies anónimos mostrando las páginas solicitadas en su propia interfaz. Se les tilda de anonimizadores porque remueven la información identificatoria de los usuarios antes de llevarlos a algún sitio. En la Figura 1.3 se muestra más claramente su funcionamiento.

Ejemplos de sitios que se especialicen en esta tarea son: `www.anonymizer.com` y `www.anonimouse.com`. Cabe destacar que es incontable la cantidad de sitios que proveen este servicio en Internet. Mayormente su acceso es libre, permitiendo que cualquier usuario pueda hacer uso de ellos. Aunque existen algunos que

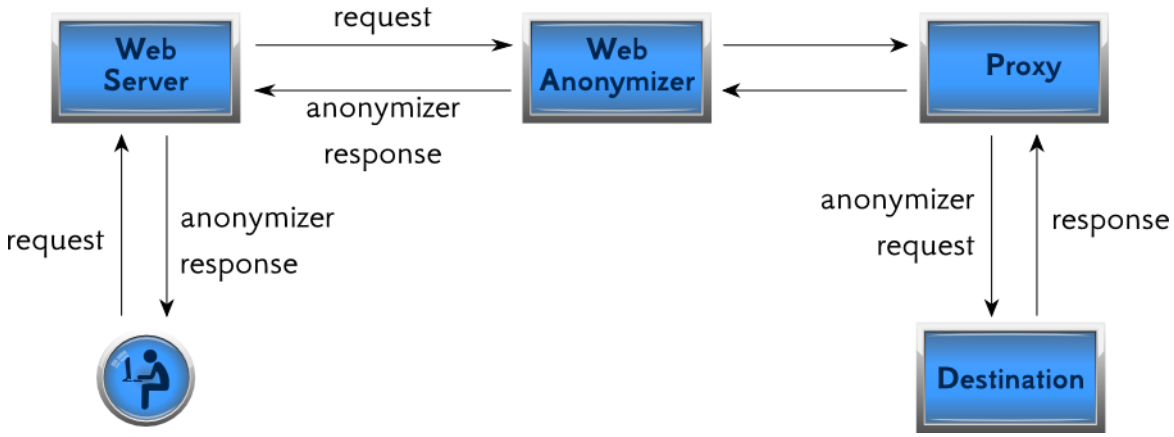

Figura 1.3: Funcionamiento de un sitio proxy

su uso es completamente libre y gratis, también se encuentran otros que precisan de un pago para obtener de ellos un uso óptimo.

Características y funcionamiento de los sitios web anonimizadores

Para la construcción de los sitios anonimizadores se utilizan diversos *scripts* proxies. Estos *scripts* son programas que una vez que obtienen la solicitud de la URL, recuperan las páginas web desde el sitio destino y les envían los resultados al usuario. Al igual que el motor de un automóvil, el *script* proxy es el corazón del sitio web proxy. Por lo tanto, el impacto del sitio proxy depende de lo fuerte y flexible que pueda resultar el *script* utilizado para su creación. Apoyándose en las siguientes características, los administradores de estos sitios basan su selección a la hora de construir sitios proxies [8].

- **Compatibilidad de la web**: Existen algunos *scripts* que no son compatibles con tecnologías tales como JavaScript[8], *pop-ups*[9], *cookies*[10], Flash Movie[11], o Hypertext Transfer Protocol Secure (HTTPS), imposibilitando la navegación por algunas páginas que se apoyen en estos recursos. En dependencia

[8]JavaScript es un lenguaje de scripting basado en objetos, utilizado para acceder a objetos en aplicaciones.

[9]Denota un elemento emergente que se utiliza generalmente dentro de terminología Web.

[10]Las cookies son fragmentos de información que se almacenan en el disco duro del visitante de una página web a través de su navegador, a petición del servidor de la página.

[11]http://flash-movie-player.softonic.com/

del uso que los usuarios finales le den al sitio, se escoge un *script* que permita o no el uso de las tecnologías antes mencionadas.

- **Rendimiento del proxy**: Debe contar con gran rapidez para procesar una solicitud. Existe una gran competencia entre estos sitios, por lo que una demora de esta índole traería malas consecuencias.
- **Fácil de usar y personalizar**: Los administradores no necesitan conocer a fondo el funcionamiento interno de estos *scripts*, pero que los mismos cuenten con funciones de configuración si es muy importante. Posibilitar administrar el tamaño de la caché, la limitación de descargas, bloquear determinados rangos de Ips, además de permitir cambiar fácilmente el diseño del sitio, el color y el contenido del sitio, son cualidades muy apreciadas en estas aplicaciones.
- **La demanda de recursos**: Cualquier solicitud de estos proxies solo consume cierta cantidad de recursos, tales como tiempo de Central Processing Unit (CPU), espacio de memoria, o de la red de ancho de banda. Teniendo en cuenta que estos sitios pueden tener más de 5.000 visitas por día, la demanda de recursos del *script* en gran medida afectará el rendimiento del sitio.
- **Continuación del desarrollo y de apoyo**: Las nuevas tecnologías de Internet, las características, los *widgets*[12], y las herramientas están en constante cambio. Para un mejor apoyo a ellos, estos *scripts* deben ponerse al día también, idealmente, deberían tener una nueva versión cada 12 meses o más a menudo. Además, deben tener una gran comunidad de usuarios con algunos foros activos que ayuden a resolver todo tipo de cuestiones.

En su mayoría estos *scripts* son muy fáciles de utilizar y de una manera sencilla generan este tipo de sitios web. Para utilizar estos sitios los mismos se valen de un campo para introducir la URL del sitio que se quiere visitar. Una vez introducida esta dirección, se adentran en la red en busca de la página que se desea visualizar y la muestran embebida en su interfaz. Si posteriormente se van siguiendo enlaces de una página a otra, se presentarán asimismo a través del anonimizador. Aunque son una de las vías más rápida y popular

[12]Es una pequeña aplicación o programa, usualmente presentado en archivos o ficheros pequeños que son ejecutados por un motor de widgets o Widget Engine

para navegar de manera anónima, también presentan una serie de inconvenientes y limitaciones en cuanto a su provecho:

- No siempre funcionan con la totalidad de los sitios, incluyéndose los servidores seguros.
- Disminuyen la velocidad de la navegación.
- Añaden banners publicitarios de sus patrocinadores a las páginas que se visitan.

El hecho de que no reciban *cookies* y desactiven los programas en Java[13], JavaScript, Flash[14] entre otros, puede resultar otro inconveniente o una ventaja en dependencia de quienes sean sus usuarios. La mayoría de las personas que utilizan estos sitios, desconocen lo maligno que pueden resultar en algunos casos. Pueden convertirse en enormes agujeros en la seguridad de la red, portales de *hackers* para el robo de datos, *spyware*, virus y gusanos, peligros de los cuales los usuarios típica y completamente se encuentran inconscientes.

1.5. Técnicas para la detección del uso de sitios anonimizadores

Debido a que son una de las mayores vulnerabilidades para asegurar el cumplimiento de las normativas de una navegación segura para redes de organizaciones o instituciones, y por tanto, uno de los eslabones más débiles para burlar los Sistemas de Filtrado de Contendido, han ido surgiendo varias técnicas que hacen hacedero detectar su uso.

1.5.1. Listas Negras

Una de las vías más practicada y conocida son las llamadas Listas Negras. Estas listas recogen las direcciones de sitios anonimizadores conocidos y mediante su consulta pueden ser bloqueados. Esta técnica se caracteriza por su bajo costo para el sistema que las implemente, ya que su utilización mediante reglas de bloqueo o de registro de búsquedas no afecta gravemente el rendimiento del sistema.

[13] Es un lenguaje de programación de alto nivel, orientado a objetos

[14] Es una aplicación en forma de estudio de animación

Los sitios anonimizadores son encontrados por los usuarios a través de diversas listas que proveen algunos sitios encargados de darles seguimiento. Estas listas se retroalimentan con URLs que son enviadas por nuevos propietarios de sitios anonimizadores, siendo renovadas de esta manera diariamente. Haciendo de esto una ventaja, se pueden tomar los datos que figuran en estas listas para crear una alimentación automática y mantener actualizadas las Listas Negras.

Para recuperar las listas de los sitios que las proveen se pueden utilizar varios métodos. Uno de ellos es mediante el comando **CURL** del sistema GNU/Linux. Este comando tiene como objetivo descargar ficheros de sitios web. Tomando por ejemplo el sitio www.mwolk.com, el cual provee una lista de sitios anonimizadores[15], la sintaxis para utilizar este comando seria [9]:

grep url \"\>\<a\ href = \" http proxy_list.html — cut -d \"-f 6 > blacklist.txt

De esta manera se obtiene en un archivo local el HTML con la lista de los proxies. Una vez recuperada, con la utilización del comando GREP, se pueden seleccionar sola las líneas que contengan las direcciones URLs de los proxies publicados recientemente, además de filtrar la URL real del resto de la línea en el código HTML. Sintácticamente tendría la siguiente forma [9]:

grep url\"\>\<a\ href=\"http proxy_list.html — cut -d \" -f 6 > blacklist.txt

Aunque constituyen un importante recurso en la tarea de detectar sitios proxies, las Listas Negras son una de las vías menos eficaces debido a que tienen que ser sitios conocidos con antelación para poder ser añadidos a las mismas. Esto trae consigo además; que no se ajuste esta técnica a la solución que se necesita, ya que no posibilita la detección en tiempo real de los sitios anonimizadores.

1.5.2. Identificación de sitios proxies mediante patrones en sus URLs

Al ser construidos los sitios anonimizadores sobre *scripts* o códigos bases comunes, hace posible que se detecten patrones que ayuden a su identificación. Uno de estos patrones es la manera en que generan las URLs. A continuación se muestran los patrones y las expresiones regulares que les corresponden a las URLs generadas por tres de los *scripts* más populares en este ámbito según el autor John Brozycki.

[15]la dirección es: http://mwolk.com/proxy_list.html

PHP Proxy

PHPProxy es un servidor Hypertext Pre-processor (PHP) basado en proxy que trabaja en variantes de Unix, así como Solaris y Windows. Hasta el momento no ha sido liberada una nueva versión desde 2004. Según estadísticas, existe un promedio de descargas diarias entre 25 y 50, lo cual da una idea de la cantidad de sitios que están implementados sobre este proyecto. Las URLs generadas por este proxy anónimo siguen el siguiente patrón [9]:

hostname/index.php?q=obfuscatedURL&hl=identifier.

Mediante la siguiente expresión regular, la cual responde al patrón antes mostrado, se pueden detectar haciendo una búsqueda en la URL los sitios anonimizadores construidos sobre el *script* PHP Proxy [9].

(index\.php\?q=). + (&hl).*

Para obtener un ejemplo que ilustre lo antes afirmado, al visitar la web `www.myspace.com` mediante el sitio `www.schoolsnooper.com`, el cual está construido bajo PHP Proxy, la URL generada tiene la siguiente forma: `http://www.schoolsnooper.com/index.php?q=d3d3Lm15c3BhY2UuY29t&hl=2ed`

CGI Proxy

CGI Proxy es otro popular *script* utilizado para crear sitios proxies. Es desarrollado en el lenguaje Perl. Estos tipos de servidores se pueden dividir en tres tipos de acuerdo con el nivel de anonimato que ofrecen [10].

- **Proxy transparente**: No ocultan la dirección IP de un cliente. Su tarea fundamental como regla general, es el almacenamiento en caché de la información y/o apoyo de acceso a Internet a varios ordenadores a través de una única conexión.
- **Proxy Anónimo**: Estos proxies CGI no muestran la IP real del usuario, sino su IP pública.
- **Alto Proxy**: No muestran la dirección IP de un cliente y no envían además ninguna variable que indique que se está utilizando un servidor proxy.

Es a menudo muy utilizado por administradores experimentados como una alternativa a Glype. Es preferido entre otros *scripts* de su tipo por su completo soporte de JavaScript, que proporciona una mejor compatibilidad con sitios populares. Los dos inconvenientes principales de CGIProxy son: en primer lugar se requiere relativamente más recursos del sistema, por lo que se necesita de un servidor más potente para ejecutarlo, en segundo lugar, se requieren más conocimientos técnicos para optimizar el entorno del servidor para un mejor funcionamiento.

Para identificar los sitios basados sobre este *script* proxy, se puede hacer uso de la siguiente expresión regular [9]:

(/browse\.php/).+/.+/.+(/b).+/

Glype

Este *scritp* es uno de los más utilizados para la creación de sitios proxies. Requiere de PHP 5 para su explotación. Entre las características que lo hacen sobresalir se encuentran [11]:

- **Plug and Play**: No necesita instalación.
- **Almacenamiento en caché del lado del servidor**: Mejora la velocidad para el usuario final, reduciendo la carga del servidor y aligerando las facturas de ancho de banda.
- **Soporte para javascript**: Permite la accesibilidad a una número mayor de sitios.
- **Bloqueo de usuarios por IP**: Con direcciones individuales o a través de más de un rango de IP; protege el sitio de usuarios malintencionados o abusivos
- **URLs exclusivas**: Ofrece mayor privacidad, una vez que expire el período de sesiones, las direcciones URL almacenadas en el historial se invalidan.
- **Plugins**: Permiten integrar rápidamente el sitio con código específico.
- **Panel de control de administración**: Para un fácil manejo y configuración

Las URLs generadas por estos sitios presentan el siguiente patrón [9]:

hostname/browse.php?u=obfuscatedURL&b=identifier

Sirviendo de apoyo la siguiente expresión regular, se pueden identificar sitios generados bajo este *script* [9]:

(browse\.php\?u=).+(&b).*

1.6. Identificación de los sitios accedidos a través de sitios anonimizadores

Como se había referido anteriormente, cuando un usuario se conecta a un sitio anonimizador, tiene que pasarle la dirección del sitio restringido que esté tratando de visitar. Como es típico, esta información es pasada por parámetro en la cadena de la URL. De esta manera, aunque el sitio anonimizador no se encuentre bloqueado, es posible detectar el sitio que realmente se quiere acceder si se realiza una búsqueda en la cadena de la URL. Muy a menudo los parámetros son ofuscados en el cliente antes de pasarlos al servidor destino, haciendo más difícil obtener esta información. Las técnicas más usadas para ofuscar estos datos es la codificación en Base 64[16] y Rotate 13[17].

Para saber el sitio al cual realmente se está accediendo se decodifica la parte de la URL que responde al identificador del sitio destino. Para ilustrar lo antes mencionado, tomando como ejemplo la dirección que se expone en la sección 1.5.2, si se decodificara la parte que le corresponde a la URL del sitio destino, en este caso sería la expresión **d3d3Lm15c3BhY2UuY29t&hl**, como resultado se obtiene la dirección www.myspace.com.

1.7. Servicios que pueden ser utilizados como proxies anónimos sin ser este su fin específico

En ocasiones algunas herramientas son usadas más allá de su propósito original, los servicios de traducción de lenguajes son un ejemplo de ello. Estas herramientas además de traducir palabra a palabra,

[16] Sistema de numeración posicional que usa 64 como base.

[17] Cifrado de sustitución

proporcionan la facilidad de traducir una página web en todo su contendido y desplegarla en el navegador web. Valiéndose de esta opción se podría tener acceso a cualquier página web que se especificara para traducir. De esta manera, estas herramientas pueden ser utilizadas como sitios anonimizadores sin ser ese el fin para el cual fueron creadas. A pesar de ser otro camino hacia la navegación anónima, presentan ciertas desventajas cuando se usan como sitios anonimizadores. En algunos casos son inhabilitadas para navegar en páginas adicionales, ejecutar JavaScript o acceder a archivos media en las páginas que retribuyen.

Bloquear los sitios que proveen de estos servicios no sería la mejor solución, ya que se estaría privando a los usuarios de su uso legítimo. Afortunadamente la mayoría de los traductores web no ofuscan o esconden la dirección de las páginas traducidas en la URL, siendo detectable la página que fue traducida mediante el monitoreo de sus URLs.

1.7.1. Google Translator

El Traductor de Google es uno de los más populares y usados en este ámbito, se caracteriza por ser muy simple para la buena lectura y comprensión del usuario. Este servicio es gratuito y fácilmente accesible mediante la barra de herramientas que posee el motor de búsqueda Google. Presenta la mayor cantidad de caracteres para traducir entre los traductores que tienen límites, además de brindar el servicio de traducción en más de 51 lenguajes.

Para utilizar esta herramienta como proxy en la sección “Traducir una página web” se introduce la URL del sitio y se especifican los lenguajes de origen y destino. Es importante destacar que este traductor no permite que los lenguajes de origen y destino sean el mismo, sumándose esto a los inconvenientes que presenta como proxy anónimo. Para visualizar mejor la estructura que presentan las URLs generadas por esta herramienta del idioma, se muestra la siguiente dirección de la traducción del ingles al español de la página www.myspace.com

http://translate.googleusercontent.com/translate_c?hl=en&ie=UTF-8&sl=es&tl=en&u=http://www.myspace.com/&rurl=translate.google.com&twu=1&usg=ALkJrhjkd9CMoL8Y5xYWxcec_tZEFZYQRw

La variable **"&sl"** indica el lenguaje origen de la página, la variable **"&tl"** indica el lenguaje al cual está siendo traducida y por último, la variable **"&ie"** muestra el carácter de codificación.

Como claramente se puede apreciar la dirección de la página traducida se muestra sin ofuscación en la URL, siendo mucho más factible determinar el sitio al cual se está accediendo.

1.7.2. Yahoo Babel Fish

Este traductor fue desarrollado por la compañía Altavista[18], lleva el nombre del animal ficticio utilizado para la traducción instantánea en la serie de Douglas Adams: Guía del autoestopista galáctico. A su vez el pescado es una referencia al relato bíblico de la ciudad de Babel y de las distintas lenguas que se dice tuvieron su surgimiento allí [11]. La tecnología de traducción de Babel Fish es ofrecida por SYSTRAN[19], una de las más antiguas empresas de traducción automática.

Si mediante este traductor se accediese al sitio www.myspace.com, la URL resultante sería:

http://es.babelfish.yahoo.com/translate_url?doit=done&tt=url&intl=1&fr=bf-home&trurl=http%3A%2F%2Fwww.myspace.com&lp=en_es&btnTrUrl=Traducir

Como se observa, la variable **"&lp"** denota los lenguajes en que se realizará la traducción. De la misma forma en que se comporta Google Translator, no encripta ni ofusca la dirección de la página para la cual se solicitan sus servicios.

1.7.3. Windows Live Translator

Este traductor online es desarrollado por Microsoft[20] utilizando la misma tecnología del Babel Fish. Este servicio es compatible con los idiomas: Alemán, Chino, Español, Francés, Japonés y, por supuesto, inglés. Este traductor utiliza la tecnología de SYSTRAN, la misma utilizada por el servicio de traducción del Altavista; el Babel Fish. Un atributo que abriga es su opción de traducir textos de informática, siendo su traducción de términos específicos más cuidadosa. El *layout* de su página es semejante al de las páginas de servicios del

[18] http://es.altavista.com
[19] http://www.systransoft.com
[20] http://www.microsoft.com

Windows Live[21] y es muy simple de utilizar. Esta herramienta del idioma sólo es compatible con las versiones 7.0 ó superiores de Internet Explorer y 2.0 ó superiores del Firefox [12]. Es capaz de traducir textos de hasta solo 500 caracteres, siendo esta una de sus principales desventajas. A todo lo antes mencionado también se le suma su capacidad de traducir páginas de Internet. La navegación en la página traducida es sincronizada, pulsando en cualquier link disponible se abrirá tanto en el idioma original como en el idioma traducido.

Para la visualización de las páginas traducidas posee cuatro tipos de visualización que son accesibles en el menú ”Modos de exhibición”:

1. **Lado a lado**: Visualización estándar, que exhibe las dos páginas simultáneamente, dispuestas lado a lado.
2. **Superior** / **Inferior**: También exhibe las dos páginas simultáneamente, dispuestas una arriba de la otra.
3. **Original con traducción enfocada**: Ese modo exhibe la página original. Para ver la traducción se ubica el cursor del *mouse* sobre el texto.
4. **Traducción con original enfocado**: La página traducida es exhibida. Para ver los textos en el idioma original se lleva el cursor del mouse al idioma en cuestión.

Visitando el sitio www.myspace.com igual que con los otros traductores, la URL resultante es:
http://www.microsofttranslator.com/bv.aspx?ref=Internal&from=&to=es&a=www.myspace.com

Como mismo se observa en los demás traductores tratados, la dirección de la página visitada se observa claramente en la URL.

1.8. Conclusiones

A lo largo de este capítulo se abordaron todos los elementos teóricos para sustentar la elaboración de la solución al problema. Partiendo de la necesidad que el producto Filpacon cuente con una herramienta que

[21] http://home.live.com/?mkt=es-es

detecte el uso de sitios anonimizadores, se analizaron otros sistemas de filtrado que cuentan con esta característica; con el objetivo de estudiarlas y reutilizarlas. Esto arrojó la necesidad de construir una herramienta a la medida de Filpacon, ya que dichas características son privativas. Partiendo del estudio de las vías existentes para detectar sitios proxies se sentaron las bases para la concepción de un prototipo funcional capaz de detectar en tiempo real el uso de dichos sitios web.

Capítulo 2

Propuesta de solución

Después del análisis de varias vías se hace necesario contar con otros elementos para hacer mucho más eficiente y robusto el proceso de reconocimiento de un sitio anonimizador en tiempo real. En el presente capítulo se expone otra técnica para este fin resultado de la presente investigación, la cual se basa en la búsqueda de determinados patrones en el código HTML de los sitios web. Además, se analizarán todos los elementos a tener en cuenta para la construcción de un primer prototipo funcional que se acerque a la solución deseada.

2.1. Parseando el código HTML

La mayoría de los sitios web utilizan plantillas prediseñadas para su presentación visual. Los motivos por los cuales se debe su uso casi generalizado, es que disminuyen la dificultad creativa o técnica para realizar diseños propios de los administradores, brindándoles a los sitios web una imagen profesional, confiable y seria, además de que validan perfectamente estándares y poseen niveles aceptables de accesibilidad. De esta manera, se han convertido en una de las vías más rápida y económica para la creación de una web. Los sitios proxies no se ven exentos de la utilización de este recurso.

Además de traer consigo todas las ventajas antes mencionadas, esto hace que se puedan identificar determinados patrones ya que la mayoría de dichas plantillas mantienen la misma esencia en todos sus diseños.

2.1.1. Patrones identificados

La mayoría de los sitios proxies una vez que devuelven la página solicitada embebida en su propia interfaz, muestran un formulario en la parte superior como el que se evidencia que en la Figura 2.1, en el cual brindan la posibilidad de volver a ingresar otro destino que se quiera visitar, además de reflejar opciones con respecto a: la visualización del contenido, la manera en que se decodificara la dirección URL, el almacenamiento o no de las *cookies*, entre otras.

Esta información es respaldada por la observación de más de 300 sitios proxies pertenecientes a diferentes dominios y países. Contando con el código fuente de las páginas tras haber realizado un párser del mismo, se pueden identificar los elementos mencionados anteriormente, sirviendo su presencia de otro factor que se agrega para determinar si un sitio entra en la clasificación de ser anonimizador o no.

2.2. Identificando traductores

Como se mencionaba en la sección 1.7, la mayoría de los traductores web no ofuscan o esconden la dirección de las páginas traducidas en sus URLs. Esta característica permite que sus URLs sean fácilmente monitoreadas mediante expresiones regulares y por tanto que sean detectables las páginas que mediante ellos se visitan.

Actualmente en el producto Filpacon existe una defensa contra el mal uso de estos servicios. Una vez que el proceso de filtrado se inicia cuando un cliente solicita un recurso de Internet mediante una petición HTTP al servidor Proxy-Cache-Squid, el cual es parte inseparable del sistema, antes de devolverle al usuario la respuesta a su solicitud de información le envía esta última a un proceso Redirector, el cual es un programa que se encarga de analizar la petición del usuario y mediante una serie de verificaciones determina si el contenido será permitido o denegado.

Debido a que la URL es accesible desde la petición el proceso de identificar las páginas a las cuales se está accediendo mediante los traductores se realiza en esta sección. En estos momentos solo se aplica este proceso a las páginas traducidas por el traductor de Google. La inclusión de otros traductores solo depende

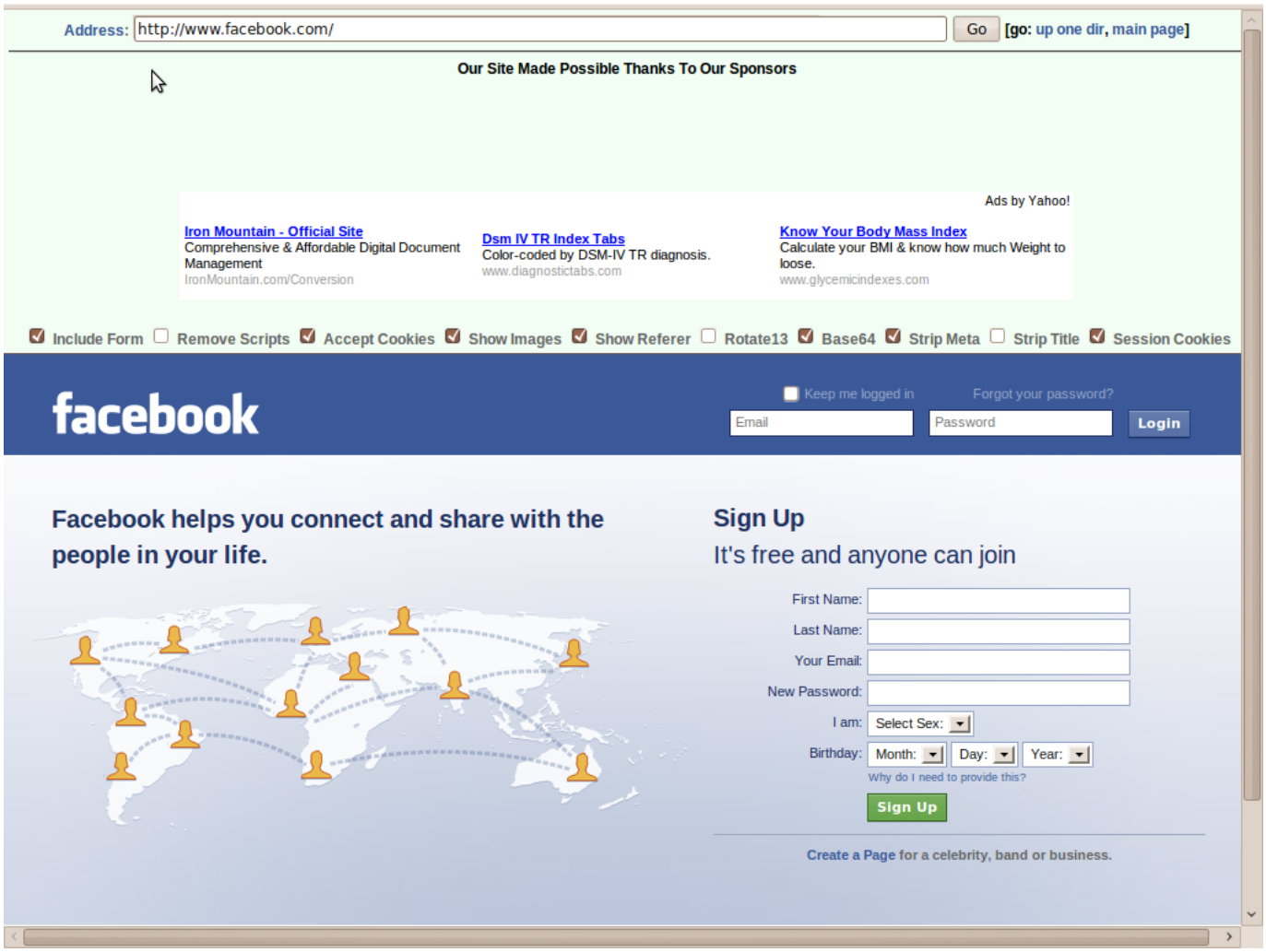

Figura 2.1: Sitio proxy

de la incorporación al sistema de las expresiones regulares que macheen con sus URLs. A pesar de que esta funcionalidad ya está contemplada en el sistema Filpacon en la sección del Redirector, en un futuro se pudiera añadir a la solución propuesta para hacerla más general y no solo encaminada a este producto.

2.3. Clasificación de los sitios. Aprendizaje Automático

Una vez obtenidos todos los elementos que tributen a la clasificación de los sitios en anonimizadores o no, se hace necesario contar con una vía para clasificarlos en dependencia de los valores que arrojen los indicadores que sean utilizados para ese fin.

Una de las ramas de la Inteligencia Artificial sirve de apoyo en este aspecto: el Aprendizaje Automático, el cual tiene como principal objetivo crear programas capaces de generalizar comportamientos a partir de una información no estructurada suministrada en forma de ejemplos. Es por lo tanto, un proceso de inducción del conocimiento [13].

Entre las ventajas que aporta el aprendizaje automático a la clasificación es que cuenta con tasas de precisión comparables a las ofrecidas por un experto humano, así como un ahorro en términos de trabajo y potencial humano, ya que no se requiere del conocimiento de un experto en un determinado tema para la construcción de un clasificador, ni en el caso en que se desee cambiar el conjunto de categorías. En general existen dos enfoques principales dentro del aprendizaje automático:

1. Aprender para poder generar un nuevo conocimiento o comportamientos para un determinado sistema.
2. Aprender para tratar de mejorar el comportamiento de un sistema.

En el primer caso se suelen utilizar técnicas basadas en razonamiento inductivo, mientras que la segunda suele estar relacionada con la utilización de técnicas analíticas. Ambos enfoques pueden emplearse también conjuntamente [13].

Dentro del aprendizaje automático, y dependiendo de si se dispone o no de datos etiquetados, se puede distinguir entre: aprendizaje supervisado y no supervisado. El aprendizaje supervisado se construye sobre

un conocimiento a priori, siendo en este caso el utilizado en este trabajo investigativo.

Se debe disponer de una colección de entrenamiento (en este caso conjuntos de datos que respondan a sitios clasificados en anónimos o no). Después de una etapa de entrenamiento, el sistema queda ajustado de tal modo que ante nuevos ejemplos, el algoritmo es capaz de clasificarlos en alguna de las categorías existentes. Cuanto mayor sea el conjunto de datos etiquetados (colección de entrenamiento) mayor será la información potencial disponible y, previsiblemente, mejor resultará el aprendizaje [14].

2.3.1. Redes Neuronales. Multilayer Perceptron

Las redes neuronales consisten normalmente en un número de elementos de procesamiento o neuronas interconectadas. Las conexiones son arreglos entre neuronas y la naturaleza de estas determina la estructura de la red. Como la fortaleza de las conexiones es ajustada o entrenada para alcanzar un comportamiento deseado, la red es gobernada por sus algoritmos de aprendizaje. Las redes neuronales pueden ser clasificadas de acuerdo a sus estructuras o algoritmos de aprendizaje [14]. De todas las existentes que cuenten con aprendizaje supervisado, Multilayer Perceptron fue la utilizada en este caso debido a que se ocupa de resolver problemas de asociación de patrones.

Multilayer Perceptron

El Multilayer Perceptron es una red neuronal artificial formada por múltiples capas, esto le permite resolver problemas que no son linealmente separables, lo cual es la principal limitación del perceptron (también llamado perceptron simple). Se caracteriza por presentar una no-linealidad en la salida, capas de neuronas ocultas y un alto grado de conectividad. Utiliza el algoritmo de retro propagación del error, que está basado en la regla de aprendizaje por corrección de error, considerada como una generalización del algoritmo de los cuadrados mínimos (LMS), utilizado en filtrado adaptivo mediante redes lineales simples. Su operación consta de dos fases, una directa y una inversa o de retroceso. En la fase directa, se ingresa el patrón de actividad en la capa de entrada de la red (vector de entrada), que recorre todas las capas subsiguientes. Se obtiene la respuesta real de la red en la capa de salida. En esta fase, los pesos sinápticos de la red permanecen fijos. En la fase inversa, los pesos sinápticos son ajustados de acuerdo con la regla de corrección del error.

Esta regla conocida como método de Levenberg-Marquardt, minimiza el cuadrado de las diferencias entre la respuesta o salida deseada y la salida real de la red [14].

El Multilayer Perceptron no extrapola bien, es decir, si la red se entrena mal o de manera insuficiente, las salidas pueden ser imprecisas. De ahí la importancia de que se realice un entrenamiento lo más óptimo posible.

2.4. Integración con Filpacon

Para realizar el proceso de filtrado como se mencionaba anteriormente, Filpacon se apoya en un programa llamado Redirector, el cual es el encargado de tomar la decisión sobre la entrega al usuario del recurso solicitado o si, en su lugar se le entregará una página de denegación en la que se le explican los motivos por los cuales se le denegó el acceso a dicho recurso. Para realizar esta operación trabaja sobre los parámetros de la solicitud, los cuales le son entregados mediante el servidor Proxy-Cache-Squid. Algunos de estos elementos son: identificador del usuario, dirección IP de origen y URL, comenzando así un proceso de interacción con la base de datos el cual dependiendo de:

- La política de navegación
- La existencia o no de la URL en la base de datos.
- Las categorías y tipo de contenido al que pertenece la URL.

Toma la decisión antes descrita [15]. El Redirector actúa sobre la petición del cliente, o sea en el modo solicitud (*request*). En el modo respuesta (*response*) no se puede accionar ya que Squid no trae contemplada en su arquitectura esta posibilidad antes de la versión 3.1 de forma nativa, derivándose que no sea hacedero el análisis del tráfico de la red ya que la versión de Squid que se utiliza en Filpacon actualmente es inferior a la que tiene soporte para lo expresado anteriormente.

El análisis del tráfico de la red es una capacidad que le provee al sistema la posibilidad de ampliar sus funcionalidades, como lo es el análisis antivirus de los paquetes que ingresan a la red, contribuyendo a que

la integridad de los recursos y la seguridad de la red se amplíe considerablemente, sumándose además, la detección en tiempo real de sitios anonimizadores.

Debido a todos los beneficios que trae consigo la incorporación al sistema de algún recurso que soporte la manipulación del tráfico HTTP, surge dicha necesidad. El protocolo Internet Content Adaptation Protocol (ICAP) es la solución a ese problema, ya que permite el filtrado de acceso en el modo solicitud y el filtrado de contenido en el modo respuesta.

2.4.1. Protocolo ICAP. GreasySpoon

El Protocolo ICAP es un protocolo de red abierto y público originado en 1999 para la redirección de contenidos con fines de filtrado y conversión. Permite el uso de antivirus, filtrado de contenidos, traducción dinámica de páginas, inserción automática de anuncios, compresión de HTML, entre otros. Los servicios basados en ICAP tienen dos posibilidades de implantación, dependiendo de si la redirección al servidor de filtrado se realiza inmediatamente después de la solicitud del cliente (modo *request*) o tras la respuesta del servidor de destino (modo *response*). Normalmente, se asocia el filtrado de acceso al modo solicitud y el filtrado de contenido al modo respuesta [13]. Existen herramientas que se encargan de implementar el protocolo ICAP. De todas las existentes se decide emplear GreasySpoon debido a todas las prestaciones que le ofrece al grupo de desarrollo del producto Filpacon.

2.4.2. GreasySpoon

Inspirado por la extensión de Firefox Greasemonkey[1], GreasySpoon permite manipular el tráfico HTTP mediante la creación de secuencias de comandos simples en varios idiomas posibles. Es una solución simple y eficaz para interceptar, filtrar y transformar el tráfico de Internet sobre la marcha[16]. En la Figura 2.2 se muestra el funcionamiento de esta herramienta.

Soporte Java / Javascript por defecto, pero puede ampliarse fácilmente con otros lenguajes de *scripting*: Ruby, Python, AWK, Groovy. Proporcionando una interfaz basada en web trata de ser lo más fácil posible,

[1] https://addons.mozilla.org/es-ES/firefox/addon/748

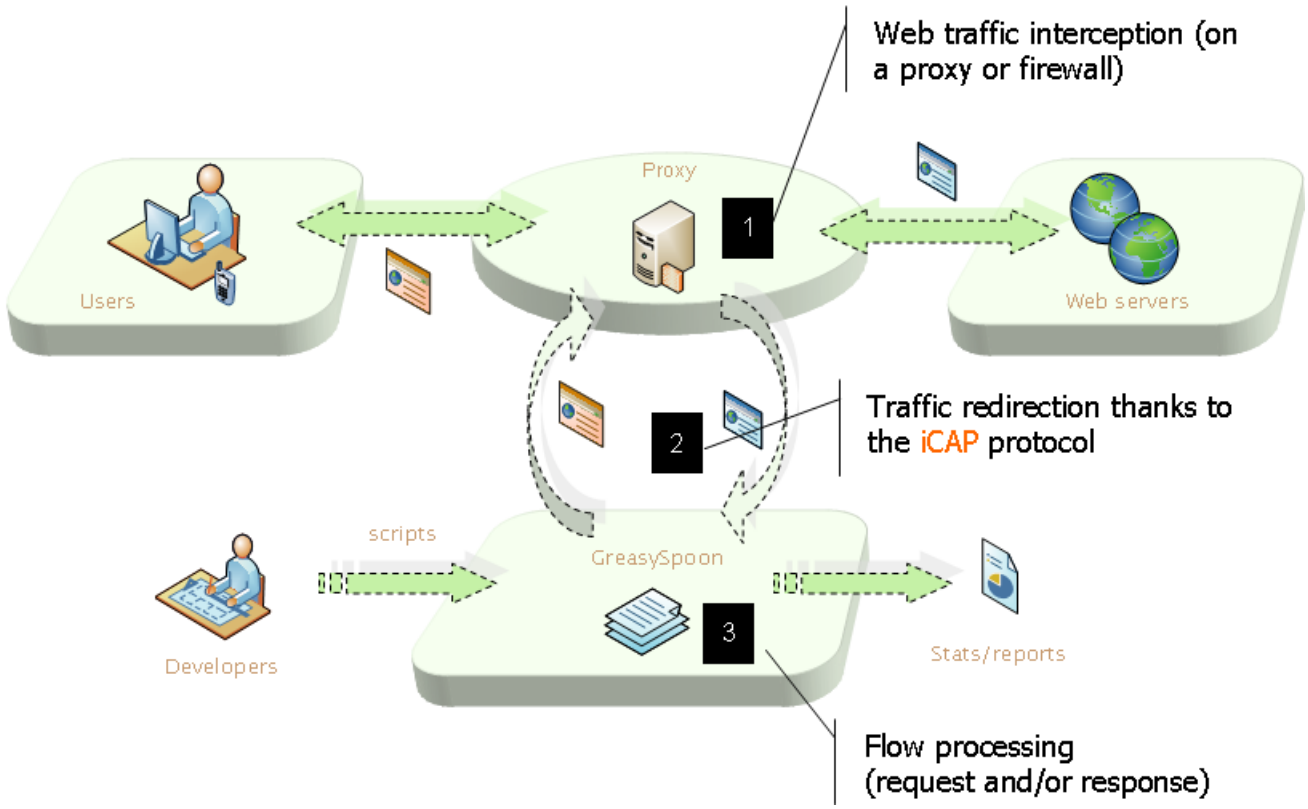

Figura 2.2: Funcionamiento de GreasySpoon

sin perder de vista en el rendimiento que lo hace utilizable desde las primeras etapas de prototipos hasta el entorno de producción.

Esta herramienta es completamente desarrollada en Java. Se puede utilizar en cualquier plataforma simple y ya ha sido probada con éxito con varios clientes del ICAP: Squid 3.0, Bluecoat ProxySG, Network Appliance NetCache entre otros más. Las principales características que presenta son:

- Facilita la construcción de servicios web en minutos.
- Crear, modificar y actualizar sus servicios de forma integral.
- Tiene soporte para varios lenguajes como: Java, JavaScript, Ruby, entre otros.
- Permite construir analizadores de contenido web.
- Es compatible con el servidor ICAP 1.0.
- No tiene ninguna limitación en cuanto al número de secuencias de comandos.
- Es independiente de la plataforma.
- Posee un motor de alto rendimiento: 1000 r / s con una latencia inferior a 20 ms en IBM xs336.

Para ejecutar esta herramienta es necesario contar con los siguientes requisitos:

- Un servidor con Java 1.6 Runtime Environment.
- Un cliente ICAP 1.0 compatible.

En la Figura 2.3 se representa el flujo interno que sigue esta herramienta para el análisis del tráfico HTTP.

2.4.3. Java

Entre los lenguajes que soporta la herramienta GreasySpoon se decidió emplear Java debido a que es un lenguaje que fue diseñado para crear software altamente fiable. Para ello proporciona numerosas comprobaciones en compilación y en tiempo de ejecución. Sus características de memoria liberan a los programadores

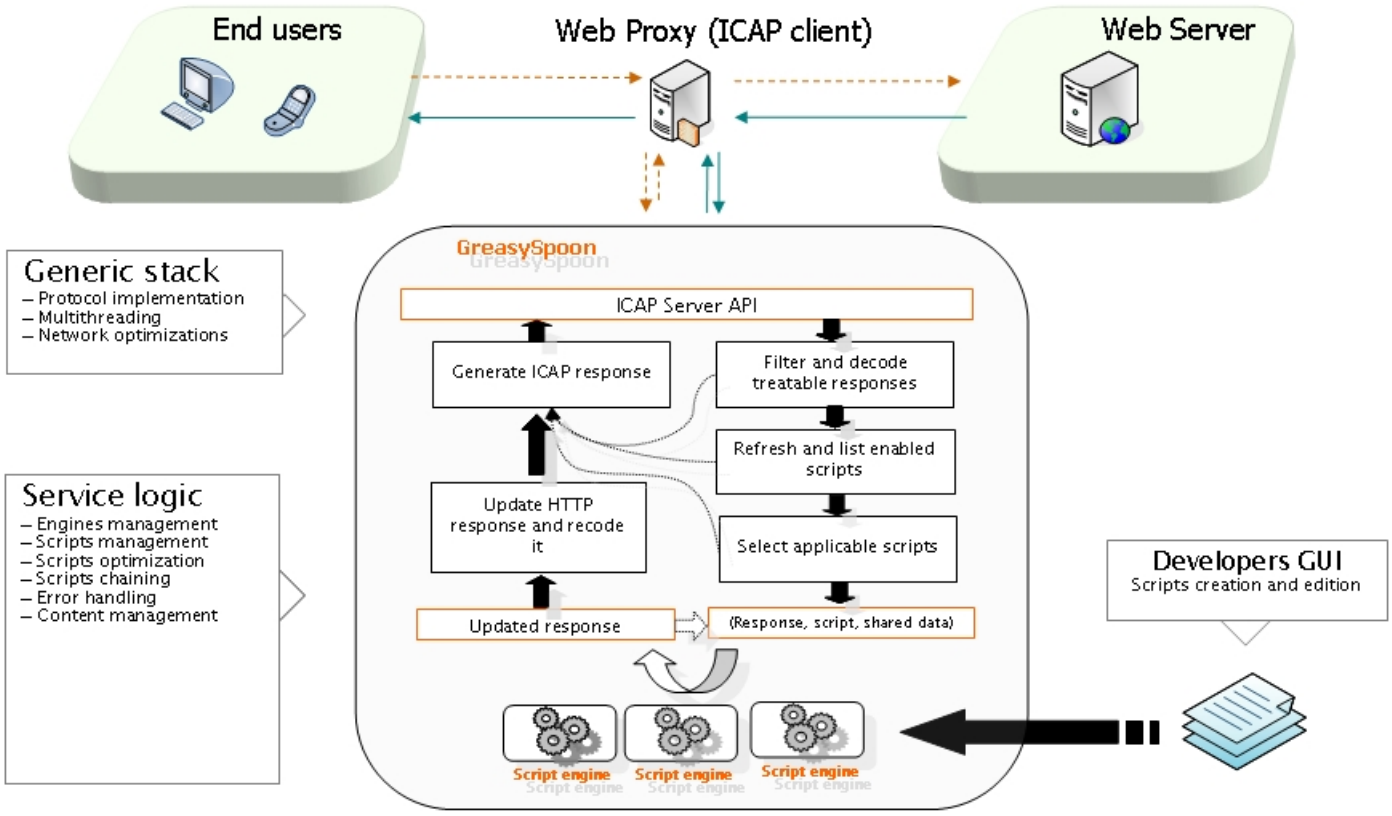

Figura 2.3: Flujo interno que sigue GreasySpoon

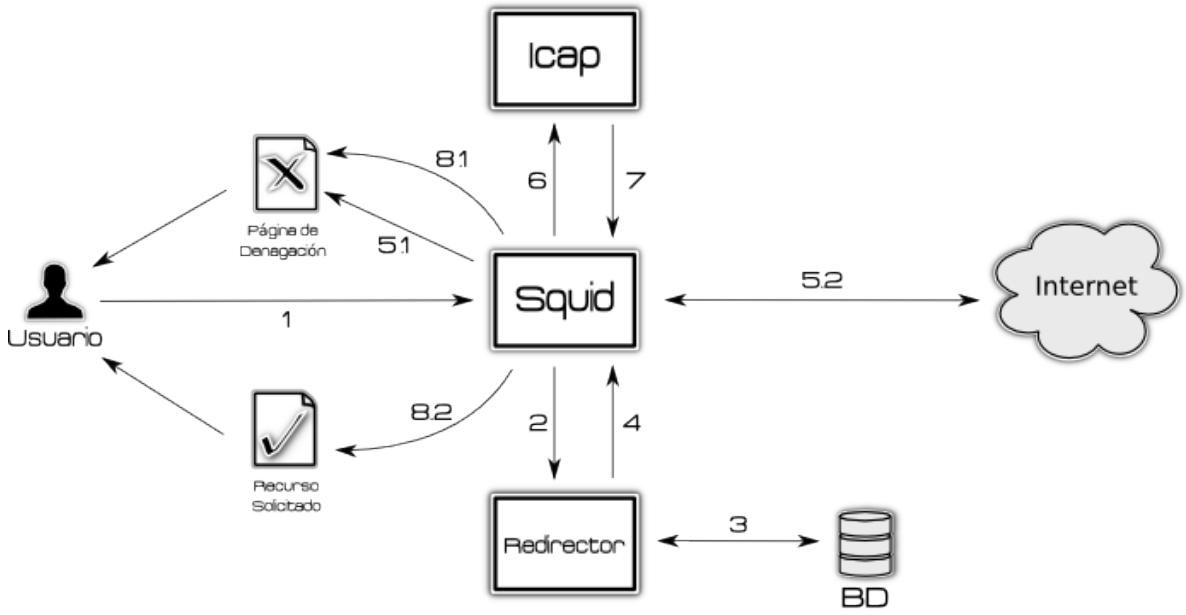

Figura 2.4: Arquitectura de Filpacon

de una familia entera de errores (la aritmética de punteros), ya que se ha prescindido por completo los punteros, y la recolección de basura elimina la necesidad de liberación explícita de memoria [17]. Además de lo antes mencionado, entre los lenguajes que soporta por defecto GreasySpoon, Java es el que más prestaciones ofrece para el trabajo con redes neuronales y el parseo del código fuente de un sitio web, bases fundamentales para la implementación de la solución que se propone en la presente investigación.

2.4.4. Arquitectura

Una vez que el protocolo ICAP se integre con el funcionamiento de Filpacon, la arquitectura que presenta actualmente este software variará en algunos aspectos. En la Figura 2.4 se muestra el escenario en el cual interactúan el servidor Proxy-Cache-Squid, el Redirector y el protocolo ICAP, también se muestra el flujo que seguirán estos componentes.

A continuación se explica el funcionamiento mostrado en la figura anterior:

1. Al servidor Proxy-Cache-Squid le llega la petición HTTP realizada por el usuario.

2. El servidor Proxy-Cache-Squid le envía la solicitud al programa Redirector para realizar el filtrado de acceso en modo solicitud .

3. El Redirector consulta en la base de datos del sistema la configuración de las políticas de navegación del usuario.

4. El Redirector le comunica al servidor Proxy-Cache-Squid si debe permitir o denegar la petición hecha por el usuario.

5. El Redirector le comunica al servidor Proxy-Cache-Squid que debe denegar la petición, este le muestra al usuario una Página de Denegación.
   5.2 El Redirector le informa al servidor Proxy-Cache-Squid que la solicitud es aceptada, dando paso al análisis sobre el modo respuesta, el cual se inicia con la recuperación del recurso solicitado en Internet.

6. El servidor Proxy-Cache-Squid le entrega al protocolo ICAP el recurso para comenzar su análisis y determinar si puede ser permitido o denegado al usuario.

7. ICAP le informa al Proxy-Cache-Squid el resultado del análisis que se realizó sobre el modo respuesta.

8. El recurso no puede ser permitido, por lo que el servidor Proxy-Cache-Squid le entrega al usuario una Página de Denegación.
   8.2 El servidor Proxy-Cache-Squid le entrega al usuario el contenido que inicialmente solicitó, debido a que el análisis concluye que el recurso es permitido.

Gracias a que el protocolo ICAP permite la centralización del análisis de la solicitud y la respuesta, se pretende que en próximas versiones de Filpacon todo el análisis que se encarga de hacer el Redirector se realice sobre dicho protocolo, provocando así que el Redirector deje de ser un componente del sistema Filpacon.

2.5. Implementación del prototipo funcional

A modo de resumen, se concreta que la solución que propone esta investigación está basada en la identificación de las URLs provenientes de sitios proxies construidos sobre los *scripts* PHP-Proxy, CGI y Glype, y

en el reconocimiento de patrones en el código HTML de las páginas.

Una vez aplicadas cada una de las técnicas referidas, se utilizará un algoritmo de clasificación automática, en este caso una red neural del tipo Multilayer Perceptron, la cual se encargará de decidir si el sitio entra en la clasificación de ser anonimizador o no.

2.5.1. Modelo de Dominio

El Modelo de Dominio (o Modelo Conceptual) es una representación visual de los conceptos u objetos del mundo real significativos para un problema o área de interés. Representa clases conceptuales del dominio del problema, conceptos del mundo real y no de los componentes de software. Identificar muchos objetos o conceptos forma parte de una investigación del problema. El lenguaje Lenguaje Unificado de Modelado (UML) contiene la notación en diagramas de estructura estática que explican gráficamente los Modelos de Dominio. El paso esencial de un análisis orientado a objetos es descomponer el problema en conceptos individuales; ya que una representación de conceptos en un dominio del problema se ilustra con un grupo de diagramas de estructura estática donde no se define ninguna operación. En estos modelos se muestran:

- Conceptos
- Asociaciones de conceptos
- Atributos de conceptos

En la Figura 2.6 se ilustra el modelo de dominio que describe el problema que se quiere resolver. Teniendo como punto de partida el modelo de dominio, surge el siguiente diagrama de clases (Figura 2.5), el cual recoge los principales elementos para la implementación del prototipo funcional.

2.5.2. Entrenamiento del algoritmo de clasificación

Para el entrenamiento del algoritmo propuesto se necesitan el código HTML y las URLs de sitios anonimizadores y no anonimizadores. Actualmente no existe una arquitectura de red neuronal que sea adecuada para una tarea específica como es el caso de la clasificación de un sitio web. Por tal razón, la selección de

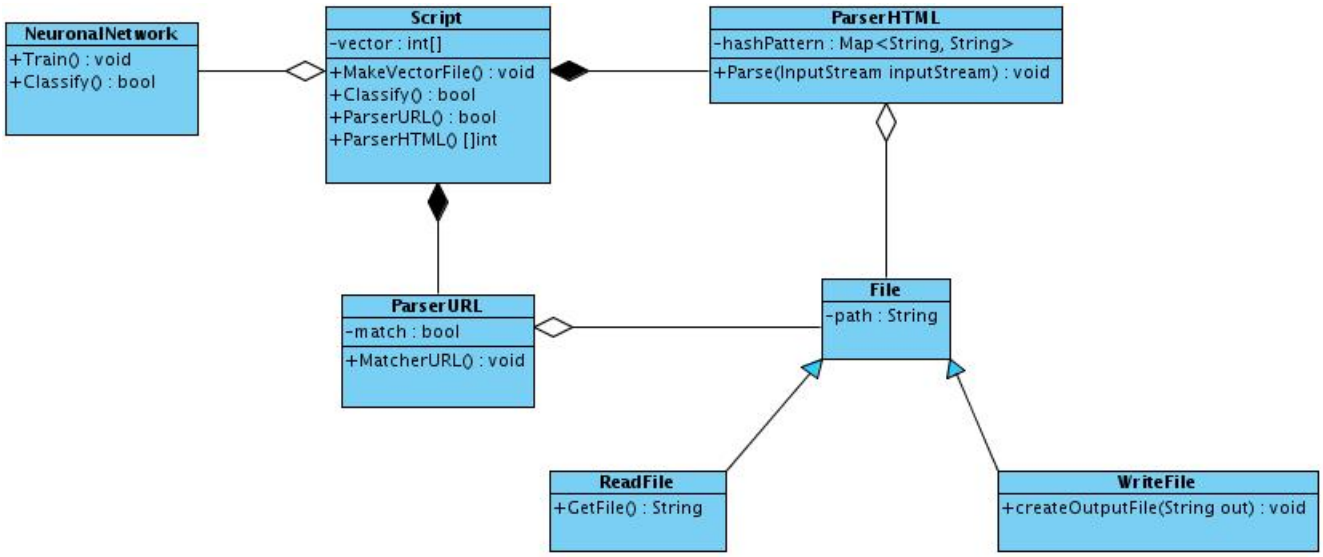

Figura 2.5: Diagrama de clases

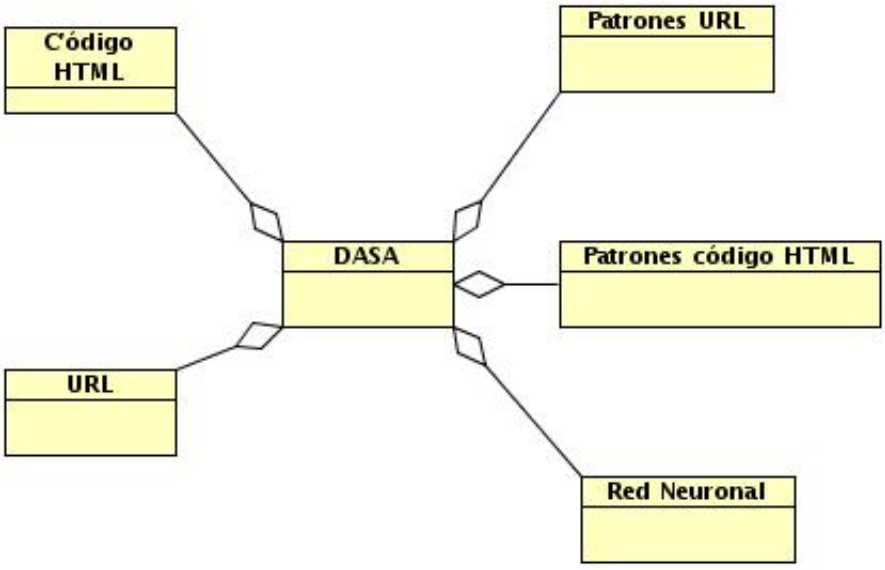

Figura 2.6: Diagrama de dominio

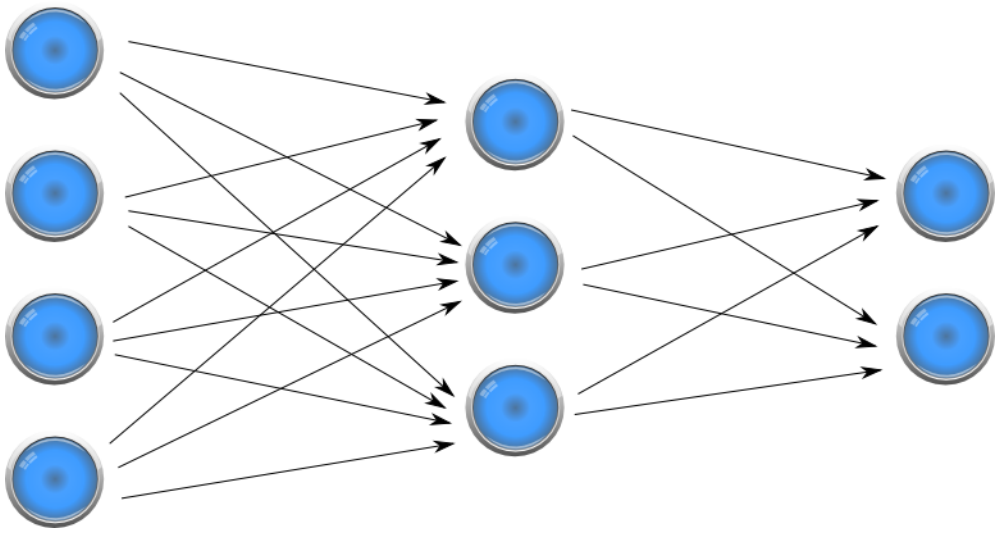

Figura 2.7: Red Neuronal primer criterio

una u otra arquitectura no deja de ser un problema de prueba y error, donde se invierte gran cantidad de tiempo probando diferentes arquitecturas y formas de entrenamiento.

Para la selección de la arquitectura predominan dos criterios. Algunos autores plantean que una red neuronal de *n* entradas y *m* salidas, se debe comenzar con una capa oculta *P*, donde $P = \frac{n}{2} + \frac{m}{2}$ como muestra la Figura 2.7, en cambio, otros autores expresan de que debe comenzar con una capa oculta $P = \texttt{MAX}(n, m) + K, dondeK \in \Re[1, 5]$ como muestra la Figura 2.8.

En el caso de esta investigación, la arquitectura utilizada quedó de la siguiente manera:

Cantidad de entradas (*n*): 16, cantidad de descriptores

Cantidad de salidas (*m*): 1, cantidad de categorías para la clasificación

Cantidad de neuronas en la capa oculta (*P*): 18. Este resultado fue obtenido utilizando uno de los criterios analizados: $P = \texttt{MAX}(16, 1) + 2$.

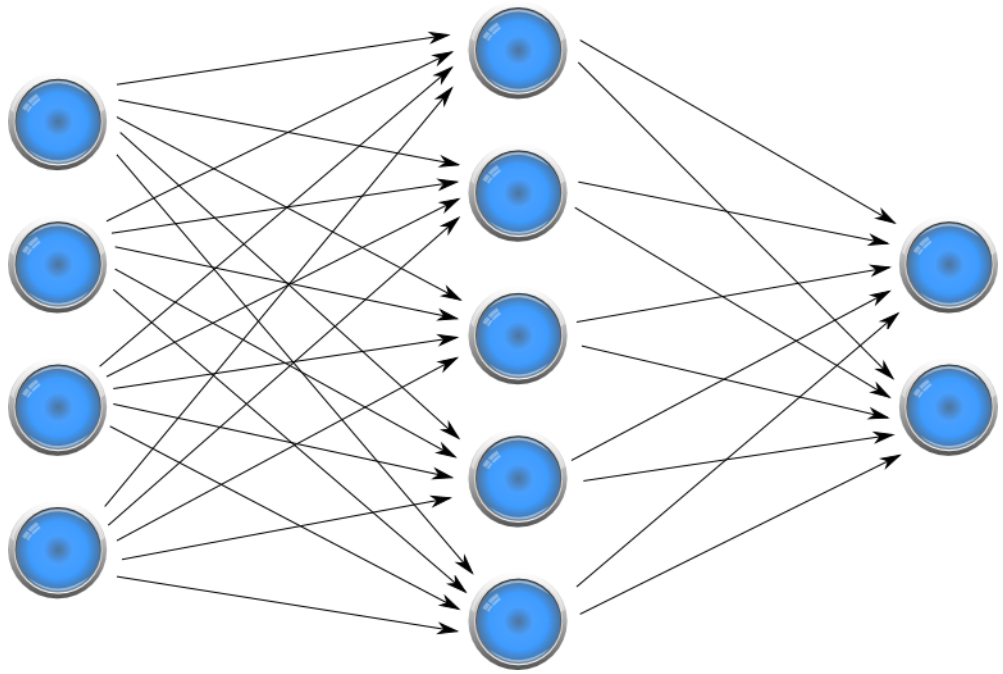

Figura 2.8: Red Neuronal segundo criterio

2.6. Probando la efectividad

Para evidenciar el grado de efectividad con que cuenta el prototipo funcional se realizó una prueba de clasificación con una muestra de 33 sitios proxies y 20 sitios no proxies. Para lograr dicho objetivo se calculó la Medida F o F1 como es llamada comúnmente. En la estadística esta es una medida que permite conocer la exactitud de una prueba. La cuenta F1 puede ser interpretada como un promedio ponderado de la precisión (*precision*) y la memoria (*recall*), donde una cuenta de F1 alcanza su mejor valor en 1 y el peor en 0. La fórmula sobre la cual se calcula es:

$$F = \frac{2 * precision * recall}{precision + recall}$$

La precisión es una medida de la exactitud que proporciona una clase especificada que ha sido predicha. Esta medida está definida como:

$$precision = \frac{tp}{tp + fp}$$

Donde **tp** y **fp** son los números de predicciones verdaderas positivas y falsas positivas para la clase considerada. En este caso **tp** sería la cantidad de sitios que son proxies y son identificados como tal, y **fp** sería la cantidad de sitios identificados como sitios proxies pero que en realidad no lo son.

La memoria es una medida de la capacidad de un modelo de predicción de seleccionar los casos de una cierta clase de un juego de datos. Comúnmente también le llaman la sensibilidad, y corresponde a la tarifa verdadera positiva. Es definida por la fórmula:

$$recall = \frac{tp}{tp + fn}$$

Donde **fn** es el número de predicciones falsas negativas para la clase considerada. En este caso serían los sitios que son sitios proxies y no son identificados como tal.

Teniendo como base los resultados que se muestran en la tabla a continuación, tras haber realizado la prueba, se calculó la Medida F obteniendo como valor final 0.94, lo que es equivalente a un 94 % de efectividad, el cual es un resultado satisfactorio.

tp	30
fn	3
fp	0

Cuadro 2.1: Resultados obtenidos

2.7. Conclusiones

Con la culminación del presente capítulo quedó evidenciada la necesidad de incorporar al sistema Filpacon de algún recurso que permitiera el análisis del tráfico HTTP para poder implantar la solución propuesta, esto trae consigo además que este producto pueda contar con otras características como lo es el análisis antivirus de todos los paquetes que circulan por la red. Tras la implementación del prototipo funcional se le hicieron pruebas de efectividad las cuales mostraron su correcto funcionamiento.

Conclusiones

Luego de analizar la posibilidad de utilizar alguna solución existente a nivel nacional e internacional para dar respuesta a la problemática planteada, se decidió investigar y construir una solución a la medida para el sistema Filpacon. Después de finalizar la presente investigación en la cual se recogen los elementos para sustentar una futura solución, se pueden plantear las siguientes conclusiones:

- El estudio del arte relacionado con la detección del uso de sitios anonimizadores arrojó la necesidad de desarrollar una herramienta a la medida de Filpacon. Haciendo posible además contar con los elementos teóricos para el desarrollo de la solución propuesta.
- El estudio y la identificación de vías para detectar el uso de sitios anonimizadores sentó las bases para la construcción de un prototipo funcional.
- Derivándose de la vinculación a Filpacon del protocolo ICAP, dicho sistema contará con más robustez y eficacia, ampliando sus utilidades.
- Con la implementación del prototipo funcional se validaron las técnicas que utilizará la solución final.

Ante los elementos anteriores puede afirmarse que se cumplieron los objetivos propuestos y se verificó la validez de la idea a defender, lo cual se materializa mediante la implementación del prototipo funcional.

Recomendaciones

- Profundizar en otras vías que permiten la navegación anónima. Además de los sitios proxies existen otros recursos que brindan la posibilidad de navegar de incógnito en Internet, constituyendo otros agujeros de seguridad para burlar el sistema Filpacon.
- Identificar los patrones que siguen las URLs de otros *scripts* proxies además de los estudiados en la presente investigación.
- Identificar más patrones en el código HTML de las páginas de los sitios anonimizadores.
- Añadir mejoras a la solución propuesta para resolver problemas de latencia que pudieran surgir al procesar un gran volumen de información, el uso de la programación paralela se recomienda para resolver este problema.
- Realizar otros entrenamientos con otras topologías para conseguir una mejor eficacia y eficiencia en la clasificación de los sitios.
- Realizar otros entrenamientos con una mayor colección de sitios.

Bibliografía

- Sahasrabudhe, Shailendra. Risks Posed Anonymous Proxies, Marzo 2008 Disponible en: http://www.expresscomputeronline.com/20080317/technology02.shtml
- Marshall, James. Proxy in a CGI Script. What it is, what it is. 2008. Disponible en: tit<http://www.jmarshall.com/tools/cgiproxy
- John Brozycki. Detecting and Preventing Anonymous Proxy Usage, 2008
- Zoica, Remus. New Threat That Can Be Used to Divert Web Traffic Through a Malicious Proxy, 2007 Disponible en: <http://www.securitysoftwarezone.com/new-threat-that-web-traffic-review350-add-
html
- June Jamrichoja Parsons Conceptos de computación: Nuevas perspectivas, 2008
- Vilma Sánchez del Castillo La publicidad en internet: Régimen jurídico de las comunicaciones electrónicas, 2009
- Universidad Tecnologica de Pereira Principales tipos de redes neuronales.
- Tutorial de redes neuronales, 2009 Disponible en: http://ohm.utp.edu.co/neuronales
- Sergio Luján Mora. Modelo de Dominio, 2008 Disponible en: http://iie.fing.edu.uy/ense/asign/desasoft/practico/hoja8/ejemplos_clase2.pdf
- ABC.es Los servicios de navegación anónima, vía de escape hacia las páginas censuradas en Internet, 2002

- Alonso,Allende. Menores en la Red. Sistemas de filtro de contenidos nocivos. Disponible en: http://www.informatica-juridica.com/trabajos/trabajosVarios.asp
- Introducción a los Proxys y como detectarlos. 2007 Contacto: rocapal@pantuflo.escet.urjc.es.
- Tor project staff. *Tor: anonymity online.* . Disponible en: http://www.torproject.org
- Anonymous. Usage Statistics for phpproxy, 2008. Disponible en: https://sourceforge.net/project/stats/?group_id=109342&ugn=phpproxy
- Anonymous. About Glype, 2008 Disponible en: http://www.glype.com
- Anonymous. How does Base 64 Encoding Work?, 12 Septiembre 2007. Disponible en: http://www.hcidata.info/base64.htm
- Joaquin Ataz Lopez Guia casi completa de BiBTeX.
- Tobias Oetiker. An acronym environment for LaTeX.
- Raul Mata Botana. Tablas en LaTeX.
- Modelo de dominio, 20 mayo 2008. Disponible en: http://iie.fing.edu.uy/ense/asign/desasoft/practico/hoja8/ejemplos_clase2.pdf

Referencia bibliográfica

[1] Dulce Carolina Córdova Cruz (2005). *Origen del Diseño de la Interfaz Gráfica* [Internet]. Disponible en: <http://www.aladdin.com/>.

[2] iPrism Web Filter (2010). *iPrism Web Filter: Simplicidad, Rendimiento y Valor*. Disponible en:<http://www.iprism.com.mx/products/iprism/web_filtering/default.html/>.

[3] iPrism Web Filter (2010). *iPrism Web Filter. Defensa en Multi-Capas Contra los anonimizadores*. Disponible en: <http://www.iprism.com.mx/products/iprism/web_filtering/internet_security/anonymizer.html>.

[4] Insight Enterprises (2010). *Sophos Web Security and Control*. Disponible en: <http://es.insight.com/content/sophos/swsc>.

[5] Insight Enterprises (2010). *Revelamiento automático de un proxy anónimo*. Disponible en: <http://www.sophos.it/security/sophoslabs/anonymizing-proxies.html>.

[6] The Onion Ring (2009). *Tor: Anonimato online*. Disponible en: <http://www.torproject.org>.

[7] June Jamrichoja Parsons. Conceptos de computación: Nuevas perspectivas, 2008.

[8] Curl (2010). *¿Curl and Libcurl?* Disponible en: <http://curl.haxx.se/>.

[9] John Brozycki. *Detecting and Preventing Anonymous Proxy Usage*. 2008.

[10] Free CGI Proxy List (2009). *What are CGI proxy servers*. Disponible en: <http://www.freecgiproxylist.com/>.

[11] Glype Proxy (2008). *Glype Proxy*. Disponible en: <http://www.glype.com>.

[12] Bajaki (2010). *Un traductor online desarrollado por Microsoft que utiliza la misma tecnología del Babel Fish.* Disponible en: <http://www.bajaki.com/download/windows-live-translator.htm>.

[13] *Internet Content Adaptation Protocol (ICAP)*. 2001.

[14] Ana Miranda Bermudez Indira Tamarit Munoz. *Propuesta del Modulo Decisor del Motor de Clasificación Inteligente de Contenidos (MOCIC).* 2009.

[15] Raul Amambay Tarajano Perez Yaisy María Montero Martinez. *Sistema Adaptativo de Filtrado de Contenidos*. 2009.

[16] Greasyspoon (2009). *Greasyspoon: Fábrica de secuencias de comandos para los servicios básicos de red*. Disponible en: <http://greasyspoon.sourceforge.net/index.html>.

[17] Departamento de Tratamiento de la Información y Codificación (2000). *Qué es Java?* Disponible en: <http://www.iec.csic.es/CRIPTONOMICON/java/quesjava.html>.

Glosario de términos

UCI Son las siglas de Universidad de las Ciencias Informáticas la cual fue creada, por el comandante en jefe de la Revolución cubana, al calor de la batalla de ideas; en el año 2002.

Filpacon Filtrado de Paquetes por Contenidos

HTML HyperText Markup Language. Lenguaje compuesto de una serie de etiquetas o marcas que permiten definir el contenido y la apariencia de las páginas Web..

URL Uniform Resource Locator. Dirección única que identifica a una página web en Internet .

HTTP Hypertext Transfer Protocol. Protocolo que define la sintaxis y la semántica que utilizan los elementos software de la arquitectura web (clientes, servidores, proxies) para comunicarse

CSRT Content Security ResponseTeam

SSL Secure Socket Layer

P2P peer-to-peer. Es una red de computadoras en la que todos o algunos aspectos de esta funcionan sin clientes ni servidores fijos.

LDAP Lightweight Directory Access Protocol. Es un protocolo a nivel de aplicación que permite el acceso a un servicio de directorio ordenado y distribuido para buscar diversa información en un entorno de red.

IP Dirección IP. Etiqueta numérica que identifica, de manera lógica y jerárquica, a una interfaz (elemento de comunicación/conexión) de un dispositivo (habitualmente una computadora) dentro

de una red que utilice el protocolo IP (Internet Protocol)

PC Computadora Personal

TOR The Onion Ring

IRC Internet Relay Chat. Es un protocolo de comunicación en tiempo real basado en texto, que permite debates entre dos o más personas.

TCP Transmission Control Protocol. Es uno de los protocolos fundamentales en Internet.

URI Uniform Resource Identifier. Es un identificador uniforme de recurso.

ICP Internet Cache Protocol

HTTPS Hypertext Transfer Protocol Secure. Es un protocolo de red basado en el protocolo HTTP, destinado a la transferencia segura de datos de hipertexto, es decir, es la versión segura de HTTP.

PHP Hypertext Pre-processor. Es un lenguaje de programación interpretado, diseñado originalmente para la creación de páginas web dinámicas.

CPU Central Processing Unit. Es el componente en un ordenador, que interpreta las instrucciones y procesa los datos contenidos en los programas de la computadora.

ICAP Internet Content Adaptation Protocol

UML Lenguaje Unificado de Modelado