UNIVERSIDAD CARLOS III DE MADRID

La Internet que no aparece en los buscadores

TRABAJO FIN DE GRADO

GRADO EN INGENIERÍA INFORMÁTICA

Curso académico: 2011/2012

Tutor: Julio César Hernández Castro

Autora: Marta Parrilla Monrocle

Índice de contenido

Índice de tablas

Índice de figuras

Índice de ilustraciones

1.- Motivación y Objetivos

El principal objetivo de este documento es realizar un estudio de cómo y qué contenidos no aparecen en Google ni en ningún otro buscador y, de esta forma, resultan inaccesibles para la mayoría del público.

Por otro lado, también se pretende investigar sobre cómo esta técnica puede utilizarse para ocultar actividades fraudulentas.

Se busca identificar las características principales de la denominada Internet Invisible, identificar las causas de su existencia, investigar las técnicas de ocultación de contenidos en la Web y aprender cómo se pueden consultar los contenidos presentes en ella.

Para dar paso a toda la información recogida sobre la Internet Invisible, primero se hablará de los buscadores, su arquitectura y las técnicas que utilizan para indexar el contenido de la Web.

2.- Estado del Arte

Aunque este trabajo está orientado al conocimiento de la Internet Invisible (una gran desconocida para muchas personas), también se hace necesario hablar de los buscadores (sus características y funcionamiento) para entender por qué hay partes de Internet que no pueden ser indexadas.

Por esta razón este apartado va a dividirse en dos temas principales: la Web y los buscadores. Para cada uno de estos temas se va a tratar tanto el origen como la situación actual.

2.1.- La Web

2.1.1.- Origen

Los orígenes de Internet se remontan a hace más de 25 años, cuando surgió como un proyecto de investigación dentro de un ámbito militar. En 1969, en plena guerra fría, el Departamento de Defensa Americano (DoD) llegó a la conclusión de que su sistema de comunicaciones era muy vulnerable, ya que tenía el riesgo de que parte del país se quedara aislado si sucediese un ataque militar sobre las arterias de comunicación.

Para sustituir a ese sistema, el DoD a través de su Agencia de Proyectos de Investigación Avanzados (ARPA), decidió estimular las redes de ordenadores mediante becas y ayudas a departamentos de informática de diversas universidades y empresas privadas. El resultado de esta investigación fue una red

experimental de cuatro nodos denominada ARPAnet (en Diciembre de 1969). Se estableció entonces la primera conexión entre los ordenadores de 3 universidades californianas y otra en UTAH (EEUU).

Desde que se inició esta conexión hasta el día de hoy, la cantidad de usuarios y conexiones ha crecido exponencialmente cada año. El punto cumbre en el crecimiento de Internet se produjo en 1989, cuando Tim Berners-Lee redacta la propuesta referenciada a ENQUIRE (donde se materializaba la realización práctica de incipientes nociones de la Web) y describía un sistema de gestión de información más elaborado.

Tim Berners-Lee utilizó un NeXTcube como primer servidor web del mundo y escribió el primer navegador web, el World Wide Web (WWW) en 1991. Ese mismo año también creo las primeras páginas web que describían el proyecto realizado. El 6 de agosto de 1991 envió un resumen del proyecto WWW al newsgroup alt.hypertext, por lo que también se considera esta fecha como el inicio de la web como un servicio disponible públicamente en Internet. En 1990 también se creó el lenguaje HTML que, junto con la WWW, son un conjunto de protocolos que permiten de forma sencilla la consulta remota de archivos de hipertexto y utiliza Internet como medio de transmisión. Lo que se conoce como páginas web.

Ha sido tal el triunfo de la WWW que frecuentemente se usan los términos “Internet” y “Web” indistintamente, ya que los usuarios piensan que se refiere a la misma cosa. La verdad es que la “Web” es solamente una parte de “Internet”. Aparte de la WWW existen otros muchos protocolos y servicios en Internet, como el envío de correo electrónico (SMTP), el acceso remoto a otros dispositivos (SSH y Telnet), la transmisión de archivos (FTP y P2P), etc. Esta parte conocida de Internet es la que se denomina Internet Visible (o superficial) y es aquella que incluye los servicios indexados por los motores de búsqueda.

Desde las pequeñas conexiones iniciales hasta el día de hoy, el número de usuarios y conexiones ha crecido exponencialmente cada año.

2.1.2.- Situación Actual

Hoy en día, prácticamente todos los usuarios de Internet están acostumbrados a iniciar el explorador y entrar a algunos de los buscadores existentes en la red para buscar cualquier tipo de información. En estos últimos años, el uso de Internet ha sido reducido prácticamente a esto y a utilizar alguna aplicación más como por ejemplo las redes sociales. Sin embargo, esto no ocurría hace unos años, cuando comenzó Internet, ya que las conexiones se realizaban exclusivamente de forma directa hacia algún sitio web cuya dirección conociésemos de antemano.

Todavía mucha gente ignora que existe un mundo desconocido fuera del Internet conocido por todos (el que se encuentra en los buscadores), y mucho menos se sabe que este submundo de Internet, al que es difícil llegar utilizando los medios habituales de conexión, incluye un mayor número de información del que utilizan la mayoría de los internautas.

Este submundo de Internet es el conocido como Internet Invisible (o Internet Profunda, Deep Web, Hidden Web y otros tantos términos) y está compuesto por la parte de Internet que no puede ser indexada por los robots de los motores de búsqueda. Las razones principales de esta falta de indexación serán abordadas a lo largo de este documento. Para visualizar mejor el concepto de Internet Invisible, se suele mostrar la imagen de un gran iceberg del que sólo vemos una pequeña parte (la que está en la superficie) mientras que en el fondo encontramos una parte mucho mayor. Esa parte oculta sumergida en el agua es la que se conoce como Internet Invisible, y es la que guarda esa información que normalmente no podemos o no sabemos recuperar.

Ilustración 1. Visualización de Internet como un iceberg

Otro ejemplo muy ilustrativo que suele mostrarse asociado al término Deep Web es el que aparece a continuación. Esta imagen está sacada del estudio de Michael K. Bergman, The Deep Web: Surfacing Hidden Value. Bergman compara la acción de buscar información en la red con la pesca en el océano: en la superficie se puede encontrar mucha información, pero en las profundidades del océano es donde se encuentra el problema, ya que las redes no consiguen alcanzar la información (y ésta posiblemente sea más valiosa).

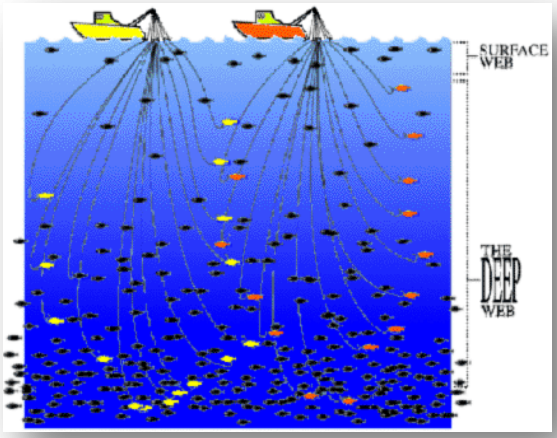

Ilustración 2. Visualización de Internet Invisible como la pesca en el océano.

En su página WorldWideWebSize.com [1], Maurice de Kunder realiza estimaciones diarias del tamaño de la web indexada. A día 4 de Junio de 2012 se obtiene como resultado 8.71 billones de páginas web indexadas. Estas estimaciones se realizan a través de los tres buscadores más importantes del momento: Google, Yahoo Search y Bing. De la suma de estas estimaciones, se resta una superposición estimada entre estos motores de búsqueda. La superposición es una sobreestimación; por lo tanto, el tamaño estimado total de la World Wide Web indexado es una subestimación.

Desde la superposición se resta en secuencia, comenzando desde uno de los tres motores de búsqueda, por lo que son posibles varios ordenamientos (y estimaciones totales). En la página se presentan dos estimaciones totales, comenzando con Yahoo (YGB) y comenzando con Google (GYB). La cifra elegida para estimar el número de páginas de la web indexada se refiere a la estimación de YGB.

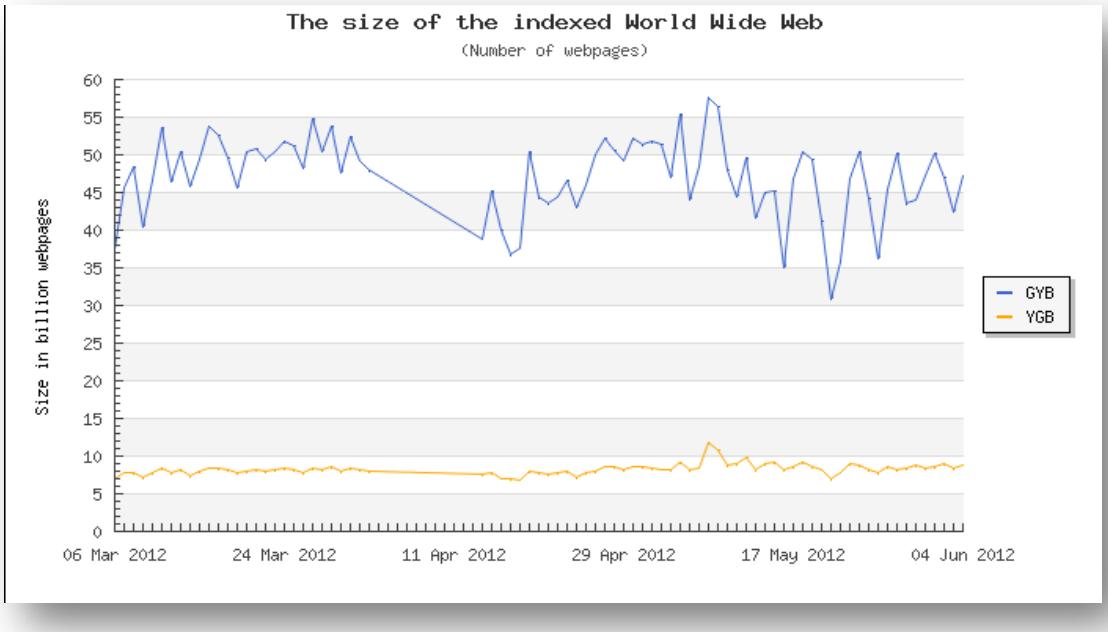

Figura 1. Tamaño de la Web indexada en los últimos 3 meses.

Fuente: http://www.worldwidewebsize.com/

Por otro lado, la empresa Netcraft nos ofrece un estudio mensual del número de dominios activos actualmente, así como la cuota del mercado de servidores en todos los dominios [2]. Tenemos que en Mayo de 2012 se recibió respuesta de 662.959.946 sitios, una disminución de 14 millones con respecto al mes anterior y la primera caída observada en 22 meses. Esta caída se ha debido a la pérdida de más de 28 millones de nombres de dominio bajo el tld .info, llevado por Softlayer. A pesar de esto, se han seguido observando un aumento de 1,2 millones de dominios nuevos.

Figura 2. Sitios totales en todos los dominios.

Fuente: **http://news.netcraft.com/archives/2012/**

2.2.- Los buscadores

2.2.1.- Origen

Con el nacimiento de la World Wide Web, surgió la necesidad de un buscador que examinase la red para que los usuarios pudiesen encontrar los contenidos que les interesasen, ya que antes únicamente podían acceder a las páginas web si conocían la dirección exacta de la misma.

El origen del primer buscador creado se remonta a 1993. Se trata de Wandex un robot de búsqueda desarrollado por Mattew Gray en el MIT que pretendía medir el tamaño de la red, pero que se amplió para poder leer direcciones URL. Al ser el primer buscador tuvo grandes problemas de infraestructura y velocidad cuando alcanzó los cientos de visitas diarias. Algunos años después, debido a la propagación de nuevos y más potentes robots de búsqueda, Wandex acabó desapareciendo.

En noviembre del mismo año, Martijn Koster presentó un nuevo buscador: Aliweb. La principal diferencia que presentaba éste sobre Wandex, es que era un motor de búsqueda de datos, no un índice. Esto le permitía añadir descripciones y palabras clave escritas para encontrar las páginas y no la dirección donde se encontraban. Además, no usaba un robot de búsqueda, por lo que no tenía el problema de la ralentización del motor por afluencia masiva de visitas. Actualmente sigue en funcionamiento y se puede encontrar en http://www.aliweb.com/ .

El primer motor de búsqueda de texto completo que apareció fue WebCrawler, en 1994. A diferencia de los dos anteriores, WebCrawler permitía a los usuarios realizar una búsqueda por palabras en cualquier página web, lo que se

convirtió en un estándar para la gran mayoría de buscadores. Fue el primero en darse a conocer ampliamente. En este mismo año también apareció Lycos.

En abril de 1994, David Filo y Jerry Yang, dos universitarios norteamericanos, crearon Yahoo! que albergaba una colección de las páginas web favoritas. El problema de Yahoo! es que comenzó siendo un directorio elaborado por personas, lo cual llevaba mucho tiempo por lo que tuvo que evolucionar incorporándole un buscador.

Con el paso de los años fueron muchos los buscadores que fueron apareciendo en la web: Altavista, Ozú, LookSmart, etc. No fue hasta el año 1998 cuando apareció Google (creado por Larry Page y Sergey Brin), el buscador más importante que podemos encontrar en la actualidad. También otros buscadores importantes hoy en día como Yandex, MSN Search (Bing), DMOZ y Baidu fueron apareciendo en los años 1997, 1998, 1998 y 1999 respectivamente.

Antes de la llegada de la WWW, ya existían motores de búsqueda para otros protocolos o usos, como es el caso de Archie (para sitios FTP) y el motor de búsqueda Verónica (para el protocolo Gopher).

2.2.2.- Situación Actual

Actualmente existe un número pequeño de buscadores realmente destacados en cuanto a su uso por los usuarios. Los principales son: Google, Bing (Microsoft), Yahoo (motor de Bing), Ask (motor de Google), Baidu (en China) y Yandex (en Rusia). En la siguiente imagen se puede ver el porcentaje de uso de estos buscadores en los diferentes países del mundo:

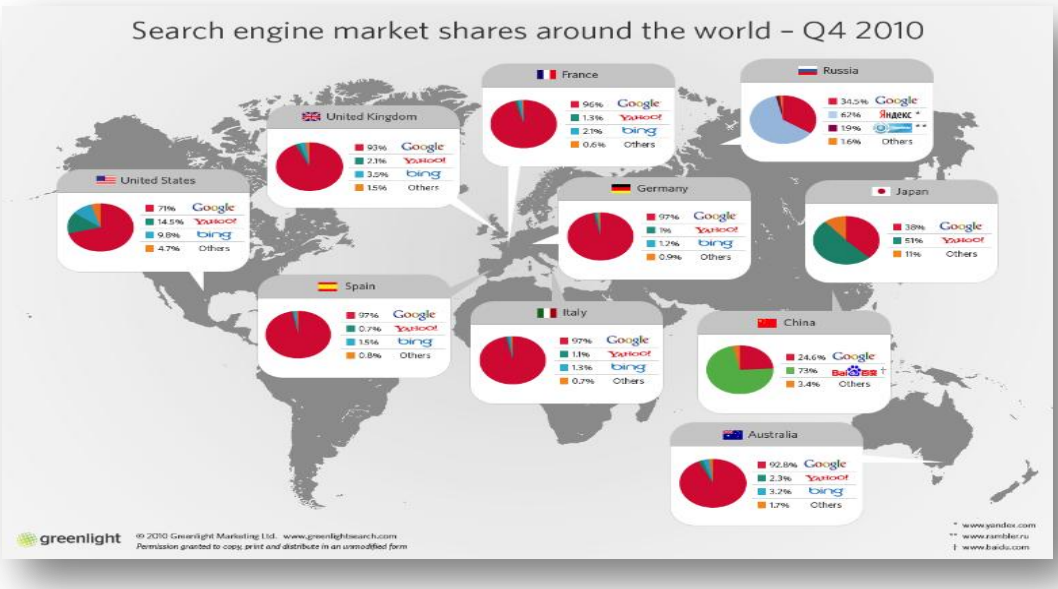

Figura 3. Porcentaje de uso de los principales buscadores.

Rojo: Google. Verde oscuro: Yahoo. Verde claro: Baidu. Azul oscuro: Bing. Azul claro: Yandex. Naranja: Otros. Fuente: http://www.onlinemarketing-trends.com/2011/02/latest-search-engine-share-of-market.html

Como se puede observar en la figura anterior, Google es el buscador más utilizado por los usuarios en la mayoría de los países del mundo, con excepción de China (Baidu), Rusia (Yandex) y Japón (Yahoo!). En España, el grado de penetración de Google es tan grande que se ha llegado a un punto en el que el 97% de las búsquedas se realizan a través de él.

Los que antes eran considerados buscadores puros (Google, MSN Search, Yahoo) en la actualidad se han convertido en compañías que ofrecen determinados servicios como noticias, correo, blogs, redes sociales, etc. El fin principal que tienen actualmente estos buscadores no es tener el menor tiempo de respuesta o devolver el mayor número de resultados, sino conseguir la fidelidad de los usuarios.

Por otra parte, en la página de Maurice de Kunder nombrada anteriormente [1], también se realiza una estimación del tamaño de los índices de cada uno de los tres buscadores mencionados anteriormente. Esta estimación podemos verla en las siguientes gráficas:

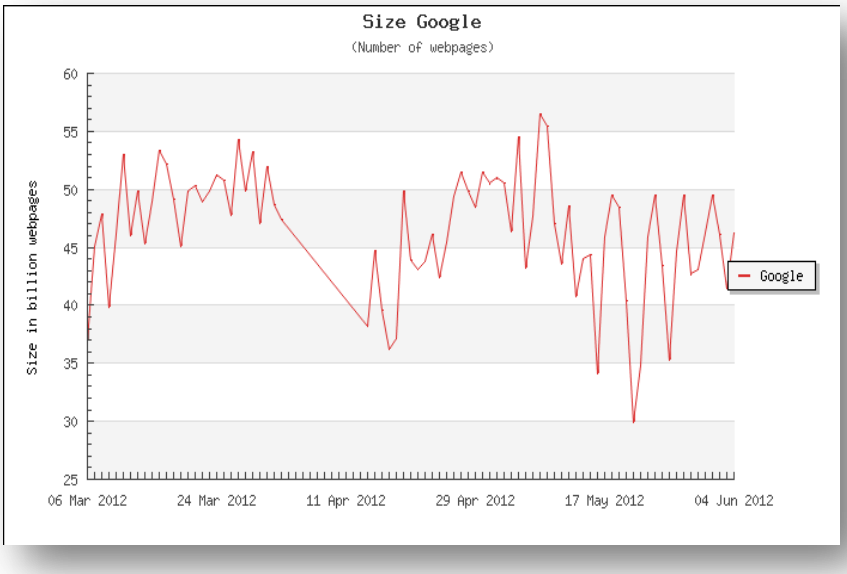

Figura 4. Tamaño de Google.

Fuente: **http://www.worldwidewebsize.com/**

El buscador Google presenta, con una gran diferencia, un tamaño de índice mucho mayor que el que tienen los otros dos buscadores. Esto se debe a que actualmente es el buscador más potente que existe ya que es el que puede abarcar el mayor contenido posible de la red.

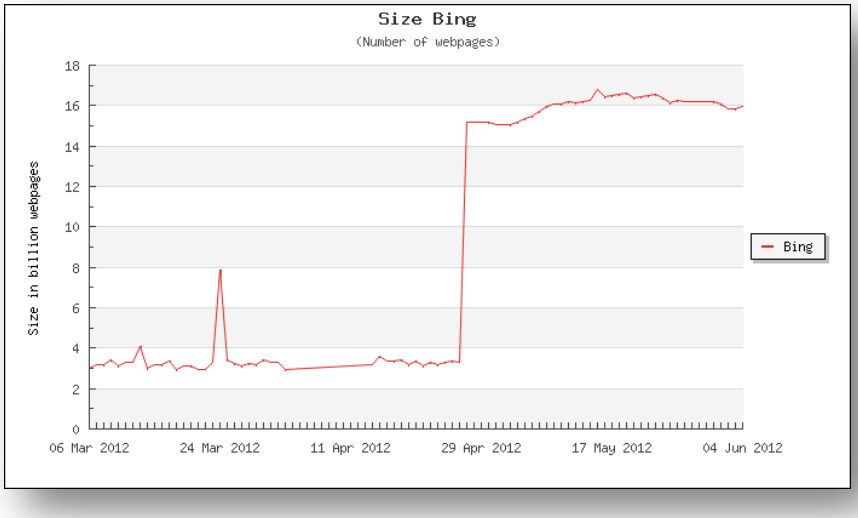

Figura 5. Tamaño de Bing.

Fuente: http://www.worldwidewebsize.com/

En cuanto al índice de Bing, se puede observar en la gráfica que en los dos últimos meses ha sufrido un considerable aumento en su tamaño.

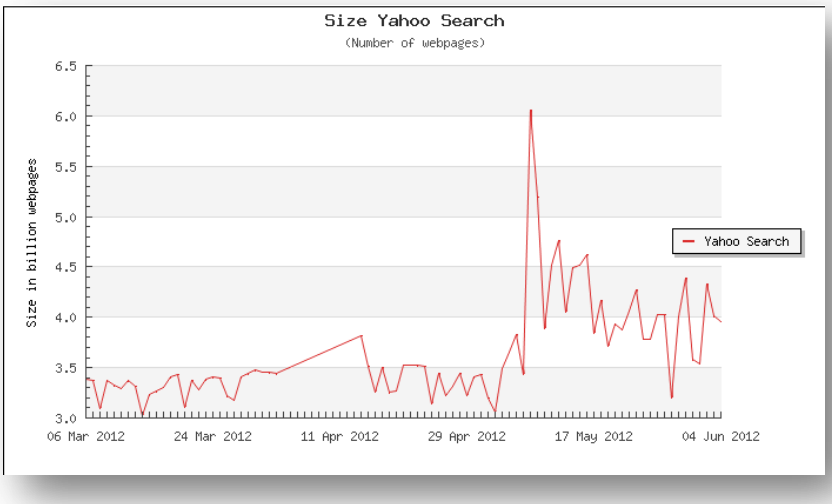

Figura 6. Tamaño de Yahoo Search.

Fuente: http://www.worldwidewebsize.com/

En el índice de Yahoo Search también se puede apreciar un notable aumento del tamaño a principios de Mayo, pero poco después volvió a descender manteniéndose en un tamaño estimado aproximadamente de 4 billones.

3.- Buscadores en Internet

3.1.- Definición de Buscador

Actualmente, Internet se ha convertido en una herramienta para la búsqueda de información rápida, para ello han surgido los buscadores, que son un motor de búsqueda que facilita a los usuarios encontrar cualquier tipo de información perteneciente a cualquier tema de interés, en cualquier parte del mundo.

Se conoce como motor de búsqueda (o buscador) al programa software o herramienta de ayuda para los usuarios, la cual les permite buscar información acerca de una temática concreta. La función que desempeña esta herramienta es buscar en bases de datos que contienen información sobre los sitios publicados en internet e indexa (crea un índice o registro de control) la sucesión de resultados obtenidos relacionados con el tema elegido o las palabras claves (keywords) ingresadas. Las palabras clave son el tema o motivo central del contenido de una página.

Los buscadores empezaron a cumplir la función de clasificación de las páginas, documentos, sitios y servidores de la red. El recorrido de las direcciones de Internet se realiza de forma diferente dependiendo del buscador, ya que éstos no emplean las mismas técnicas de búsqueda, por lo tanto cada buscador tiene una visión de la red distinta a la del resto.

3.1.1.- Clasificación

Los buscadores se clasifican en tres tipos:

3.1.1.1.- Buscadores Automáticos

Son aquellos que partiendo de cierta información conferida en lenguaje natural o en alguna otra especificación puede deducir y recuperar la información que el usuario está buscando. Su objetivo principal es encontrar los documentos que incluyan las palabras clave que el usuario introduce en la búsqueda y localizar las páginas web que mejor se adapten a dichas palabras.

Estos buscadores se componen de tres partes:

- **Robots:** Son programas que continuamente buscan información por todos los servidores web y construyen un índice de lo encontrado.

- **Bases de datos:** Contienen todos los URL encontrados, y asociados a ellos, la información relativa sobre sus contenidos:

    - Título

- Parte de texto
- Hiperenlaces
- Descriptores (palabras claves).
- etc.

Se actualizan frecuentemente por los robots que añaden nuevas páginas o referencias, actualizan las que han cambiado y borran las que ya no existen.

- **Motor de búsqueda:** Es la parte que vemos cuando realizamos la búsqueda. Cuando introducimos una petición de búsqueda, el motor la coteja con la base de datos y nos devuelve un listado ordenado con las coincidencias obtenidas. Ese listado aparece ordenado atendiendo a la relevancia de la consulta.

El funcionamiento de este tipo de buscadores consiste en lo siguiente: cuando nos conectamos con algún buscador, encontramos una página que presenta un formulario en el que tendremos que definir nuestra búsqueda y las opciones de la misma. Una vez rellenado, enviado y tras unos segundos, el buscador nos devolverá un listado de enlaces web que contienen información sobre nuestra búsqueda. Por lo tanto, tenemos dos áreas según su finalidad:

1. Formular la búsqueda y enviarla.
2. Listado de los resultados obtenidos tras la búsqueda ordenado según su similitud con las palabras clave introducidas.

3.1.1.2.- Buscadores Temáticos

Estos buscadores constituyen una guía jerárquica de directorios que va desde los temas más generales a los más específicos. Enumeran lugares (URLs), los clasifican en categorías e incorporan comentarios identificativos sobre ellos. Su objetivo es encontrar los documentos que atiendan a la temática seleccionada por el usuario.

Dichos buscadores se encuentran compuestos por dos partes:

- La base de datos, construida por los URLs remitidos.
- Una estructura jerárquica que facilita la consulta base.

El funcionamiento de estos buscadores es el siguiente: cuando nos conectamos con un buscador nos encontramos con una página que presenta una estructura jerárquica con diferentes temas. Aparece un grupo de temas generales y, según vamos seleccionando uno u otro, nos van apareciendo otro grupo de temas dependientes del anterior y cada vez más específico. Así podemos seguir hasta que encontremos el tema que andábamos buscando o finalicen las categorías implantadas por el autor del buscador.

3.1.1.3- Buscadores Especializados

Los buscadores especializados son muy parecidos a los buscadores temáticos aunque la diferencia que presentan con aquéllos es que éstos sólo abordan algún área en concreto. Además, también pueden contener buscadores automáticos. Suelen ser grandes recopilaciones del conjunto de recursos sobre un tema particular.

3.2.- Clases de Buscadores

En la actualidad existen diversos términos para nombrar a los mecanismos de rastreo, indización, recuperación y organización de documentos en la web, lo que puede dar lugar a confundir al usuario común. Cada mecanismo de búsqueda funciona y tiene un propósito y alcance diferentes, aunque con el paso del tiempo se están combinando varios mecanismos de búsqueda que dan lugar a sistemas híbridos, que pueden dificultar la comprensión del funcionamiento interno de dichos mecanismos. Otra dificultad añadida es el incremento del número de mecanismos de los que disponemos, lo que hace más necesario clasificarlos y diferenciarlos.

3.2.1.- Buscadores Jerárquicos (Arañas o Spiders)

Esta clase de buscador, conocido comúnmente como motor de búsqueda, recorre las páginas recopilando información sobre los contenidos de las mismas. Cuando los usuarios buscan una información, el buscador consulta con su software en su base de datos con la información que han recopilado de las páginas y se la presentan ordenados según su relevancia. Se puede almacenar desde las páginas de entrada hasta todas las páginas que componen una página web, esto depende de la configuración del buscador y la consideración de importancia que tenga dicha página para él.

Los resultados obtenidos al realizar una búsqueda constituirán una serie de páginas que contengan las palabras que componen la consulta que se ha realizado.

Cada cierto tiempo, el software revisa las webs indexadas para actualizar los contenidos de su base de datos, por lo que en alguna ocasión los resultados de la búsqueda no están actualizados, de forma que la información o la página no existe.

Los buscadores jerárquicos tienen una colección de programas simples y potentes con diferentes cometidos divididos en tres partes: los programas que exploran la red (*spiders)*, los que construyen la base de datos y los que emplea el usuario, el programa que explota la base de datos.

Una de las formas existentes para aparecer en las primeras páginas de resultados en los motores de búsqueda es mediante pago. Los principales buscadores delimitan estos resultados e informan al usuario que se trata de

resultados esponsorizados o patrocinados. Esta técnica se ha adoptado recientemente para poder seguir ofreciendo a los usuarios el servicio de forma gratuita.

Algunos ejemplos de este tipo de buscador son: Google, Altavista, Hotbot y Lycos.

3.2.2.- Directorios

Los directorios constituyen una tecnología barata en la que no se requieren muchos recursos informáticos. Por el contrario, requiere más soporte humano y mantenimiento. Es una tecnología ampliamente utilizada por la cantidad de programas scripts existentes en el mercado.

Son completamente distintos a los spiders, ya que los directorios utilizan algoritmos más sencillos que presentan la información sobre las webs registradas como una colección de directorios. Sólo registran algunos datos de las páginas, como pueden ser el título y la descripción introducida en el instante en que se registra el sitio en el directorio, no recorren los sitios webs para almacenar su contenido.

Los resultados obtenidos tras la búsqueda están determinados por la información que se haya proporcionado al directorio al realizar el registro de la web. Sin embargo, a diferencia de los motores, los directorios son revisados por humanos, los cuales se encargan de clasificarlos en las distintas categorías, por lo que resulta bastante más sencillo encontrar páginas que satisfagan nuestro interés de búsqueda. Más que buscar información sobre contenidos de la página, los resultados serán expuestos haciendo referencia a los contenidos y temática del sitio.

Algunos ejemplos de directorios son: Yahoo! y Terra. Ahora, ambos utilizan tecnología de búsqueda jerárquica, y Yahoo! conserva su directorio. Buscar Portal, es un directorio, así como la mayoría de motores de búsqueda hispanos.

3.2.3.- Sistemas Mixtos (Buscador - Directorio)

Los sistemas mixtos son una mezcla entre buscador jerárquico y directorio. Además de presentar características de los primeros, tienen las páginas web registradas en catálogos sobre distintos contenidos (sociedad, cultura, informática), que a su vez se encuentran divididos en distintas subsecciones.

Algunos ejemplos de este tipo de sistemas son: Excite, Voila e Infoseek. En la actualidad, los sistemas tienden hacia métodos mixtos como ha ocurrido en el caso de Altavista, aunque también pretenden parecerse a Google.

3.2.4.- Metabuscadores

Son aquellos que permiten realizar diversas búsquedas en motores seleccionados respetando el formato original de los buscadores. Su función consiste en realizar búsquedas en otros lugares, analizan los resultados de la página y muestran sus propios resultados atendiendo a un orden establecido por el sistema estructural del metabuscador. Carecen de base de datos propia.

Es necesario pedir permiso para poder utilizar los servicios gratuitos de un buscador de esta forma por el siguiente motivo: los buscadores ponen dinero para que el servicio pueda operar, el metabuscador usa estos contenidos y no da nada a cambio. Por esta razón, al omitir la publicidad, los buscadores no obtienen ingresos, sólo gastos y pérdida de usuarios que utilicen su servicio.

La ventaja principal de los metabuscadores es que aumentan de manera patente el ámbito de las búsquedas que realizamos, otorgando una mayor cantidad de resultados. Ya que cada buscador usa una estrategia distinta para recoger la información y ordenar los resultados de las búsquedas, lo que significa que las páginas que tienen mayor relevancia en un buscador no tienen por qué ser las mismas que las del resto de buscadores, obtenemos diferentes puntos de vista al recoger los documentos relevantes devueltos al realizar una búsqueda.

Por el contrario, como desventaja principal podemos indicar que los metabuscadores no distinguen entre los distintos tipos de sintaxis que emplea cada buscador, por lo que al buscar información específica es mejor emplear buscadores de los que conozcamos su sintaxis. Además, como los metabuscadores buscan la información en varios buscadores, la obtención de resultados puede ser más lenta que en los buscadores normales.

3.2.5.- FFA – Enlaces gratuitos para todos

FFA es el acrónimo del inglés “Free For All”. Se trata de pequeños directorios en los que cualquier usuario puede inscribir su página durante un tiempo limitado. Estos enlaces no son permanentes.

3.2.6.- Buscadores Verticales

Los buscadores verticales son buscadores que están especializados en un área en concreto, lo que les permite analizar la información con mayor profundidad, disponer de resultados más actualizados y ofrecer al usuario herramientas de búsqueda avanzadas. Una de las cualidades importantes de estos buscadores es que emplean índices especializados para acceder a la información de una forma más específica y fácil.

Estos buscadores envían sus robots a un número limitado de webs sobre un tema concreto, lo que hace que tanto la obtención de la información como la creación del índice sean más especializadas en el sector del que se trata. Como

utiliza un número de fuentes más reducido que un buscador genérico, los buscadores verticales actualizan su información con mayor frecuencia.

Hay diferentes tipos de buscadores, algunos de ellos están especializados en una rama de una ciencia y algunos abarcan todo tipo de materias.

Algunos ejemplos de este tipo de buscador son: Nestoria, Trovit, Wolfram Alpha e Interred.

3.2.7.- Buscadores de Portal

Los buscadores de portal son aquellos que buscan información únicamente en su portal o sitio web. Pueden considerarse como un directorio. Están basados en expresiones regulares y consultas SQL.

3.2.8.- Guías

Existen especialistas y entidades académicas que se dedican a elaborar y mantener páginas concentradoras de recursos web divididos por áreas de especialidad, en forma de directorios anotados o guías temáticas en las que se pueden encontrar recursos que no son recuperables por los buscadores comunes. Estas guías temáticas normalmente presentan un grado de calidad alto, ya que pueden poner en compromiso el prestigio de los autores e instituciones involucradas. Los recursos se seleccionan de manera cuidadosa y son actualizados de forma frecuente. Hay veces que las instituciones se asocian para formar *web rings* (circuitos) y elaborar de manera conjunta estas guías (como hacen *The WWW Virtual Library*).

Además, estas guías pueden incluir mecanismos de búsqueda en sus páginas o en la web en general.

3.2.9.- Tutoriales

En muchas ocasiones tenemos la necesidad de utilizar herramientas de búsqueda sobre las que debemos conocer su funcionamiento, qué estrategias seguir y cómo elegir los mejores instrumentos atendiendo a la necesidad de búsqueda que poseamos.

Existen tutoriales como *How to Choose a Search Engine or Directory* (de la Universidad de Albany), *SearchAbility* y *A Collection of Special Search Engines* (de la Universidad de Leiden), que orientan a los usuarios en el mundo tanto de los recursos especializados en la web como de la maquinarias que permiten su localización.

3.2.10.- Software Especializado

Hay unos programas que funcionan junto con los navegadores web y les añaden funcionalidades, denominados agentes auxiliares. Una de las funcionalidades añadidas es la de manejar conceptos en vez de palabras para recuperar información de la web. Hay otros agentes que residen en el cliente web y permiten realizar búsquedas simultáneas en varios buscadores, eliminar enlaces muertos (*dead links*), refinar los resultados de búsquedas o acceder a algunas zonas de la web invisible. Algunos ejemplos del primer tipo de agentes son: *Flyswat* y *Zapper*, mientras que un ejemplo del segundo tipo de agente es *Copernic*.

3.3.- Diferencia entre Motor de Búsqueda y Directorio

Con frecuencia se utilizan de forma indistinta los términos motor de búsqueda y directorio web, por lo que mucha gente piensa que son lo mismo cuando esta afirmación no es correcta. Como hemos visto anteriormente, los motores de búsqueda utilizan arañas (spiders) que rastrean las páginas web para indexarlas y formar un listado de sitios web, mientras que los directorios web organizan los sitios según su temática y dicha organización es llevada a cabo por personas, no por software.

Otra gran diferencia que existe entre estos dos términos es que los motores de búsqueda, al utilizar spiders que rastrean la web para indexar un gran número de páginas, tienen una base de datos bastante extensa. Sin embargo, los directorios web, al ser organizados por personas, poseen una base de datos bastante más pequeña.

Los directorios web nacieron antes que los motores de búsqueda y el primero de ellos fue Yahoo!.

Los motores de búsqueda y directorios web también se diferencian en la forma que tienen de construir la base de datos y en su estructura. En el caso de los primeros, la base de datos se relaciona mediante la búsqueda de palabras clave, mientras que en los segundos se relacionan los temas con las direcciones.

A la hora de realizar una búsqueda, tenemos más posibilidades de encontrar más resultados en un motor de búsqueda ya que, como hemos dicho anteriormente, este posee una base de datos más amplia. Sin embargo, dependiendo del tipo de búsqueda que queramos realizar nos convendrá más uno u otro. Si lo que queremos es buscar información sobre un tema genérico, es mejor usar un directorio web ya que, al estar organizado por temas, podremos encontrar diverso contenido relacionado con el tema buscado. Por el contrario, si queremos realizar una búsqueda más concreta, será más eficaz usar un motor de búsqueda, ya que tiene una base de datos más amplia y puede buscar directamente el término que queramos.

Una de las ventajas que presentan los directorios web frente a los motores de búsqueda es que en ellos siempre vamos a encontrar información relacionada

con lo que estemos buscando. Esto se debe a que, en los motores de búsqueda, como buscamos por palabras clave, podemos encontrar documentos que contengan dichas palabras, pero que no tengan la información que estemos buscando. A pesar de ello, los motores de búsqueda ofrecen herramientas para afinar las búsquedas y suprimir, dentro de lo posible, los resultados innecesarios. Esto se conoce como *búsqueda avanzada*.

La siguiente tabla resume brevemente la naturaleza de estos dos tipos de buscadores:

MOTOR DE BÚSQUEDA	DIRECTORIO WEB
Poseen bases de datos más extensas y se actualizan con mayor frecuencia.	Tienen bases de datos más pequeñas y, al intervenir el factor humano en su desarrollo son más elaboradas pero están menos actualizadas.
No tienen el contenido organizado, recogen la información de la web de forma automática y periódica.	Al ser organizados de forma manual, cuando recopilan la información la colocan por temas y categorías en sus índices.
No realizan las búsquedas en Internet "en vivo", sino en las copias de las páginas que almacenan en sus índices.	No realizan las búsquedas en Internet "en vivo", almacenan los datos de los sitios y ofrecen enlace a éstos.
Al contener más información, hay que realizar una gran explotación de sus opciones de búsqueda, por lo que son más difíciles que los directorios.	Como están organizados por temas, son más fáciles de usar, ya que simplemente hay que ubicar la búsqueda en un tema determinado.
Se utilizan para buscar información más concreta, especializada, escasa, actualizada o incorporada en páginas personales.	Son mejores para buscar información general, porque nos da como resultados páginas principales relacionadas con el tema de búsqueda.

Tabla 1. Características principales de los motores de búsqueda y directorios.

3.4.- Arquitectura de un Motor de Búsqueda

La arquitectura de un motor de búsqueda debe ser un software escalable como para soportar millones de consultas diarias y tiene que manejar grandes cantidades de datos, estableciendo un tiempo de respuesta aceptable para las búsquedas realizadas por los usuarios. Por otra parte, debe permitir la

actualización frecuente de los datos internos del sistema con los recopilados de la red.

Los motores de búsqueda están compuestos por los siguientes elementos:

- **Robot (o crawler):** Se encarga de recorrer la estructura de enlaces de la Web (estructura con forma de telaraña), mediante las listas de URLs usadas como punto de partida para el recorrido recursivo de los documentos. Los robots utilizan algoritmos para seleccionar los enlaces a seguir, determinar la frecuencia de las visitas, etc. El rastreo de la red se puede realizar de dos formas diferentes: **breadth-first**, cobertura amplia pero no profunda; y **depth-first**, cobertura amplia y profunda. Además, junto con los robots generales dedicados a descubrir recursos en la red, existen otro tipo de robots:
    - **Knowbots:** programas para localizar referencias hipertexto dirigidas hacia un documento, servidor en particular. Evalúan el impacto de las aportaciones que engrosan las áreas de conocimiento presentes en la red.
    - **Wanderes** (vagabundos): son los encargados de realizar estadísticas.
    - **Worms** (gusanos): se encargan de duplicar los directorios FTP con el fin de incrementar su utilidad a un mayor número de usuarios.
    - **WebAnts** (hormigas): conjunto de robots físicamente alejados que cooperan para la consecución de objetivos determinados.
- **Indexador:** Su misión es recibir las páginas recolectadas por el robot, extraer una representación interna de la misma y volcarla en forma de índice en una base de datos. Existen varias técnicas para extraer la información del documento (almacenar los primeros párrafos, los títulos HTML, etc.) pero las más complejas son:
    - **Extracción avanzada de vocabulario de términos:**
        - **Listas de stop (o palabras vacías):** Se trata de listas de palabras muy frecuentes que no aportan ningún significado y que no deben aparecer en el vocabulario (artículos, preposiciones, etc.).
        - **Extracción de raíces:** como su propio nombre indica, trata de recoger las raíces de las palabras para así conseguir un término único que represente a palabras parecidas (utilizada en el caso de plurales, tiempos verbales, etc.).
    - **Medidas de calidad según la frecuencia de aparición** de cada palabra en cada documento.

- **Repositorio:** Archivo donde se almacena la información útil recopilada por el robot para mostrar la salida de información que ayudará al usuario a identificar los datos de los diferentes ítems que forman el conjunto de datos como respuesta a su solicitud. La forma, estructura y datos que tiene este componente depende de la implementación específica del motor de búsqueda.

- **Servidor Web:** Es el componente encargado de establecer la comunicación con el usuario a través del protocolo http, recibiendo consultas de éste y mandándole el resultado de las mismas a través del mismo protocolo.

Por otra parte, atendiendo a la comunicación con los módulos internos del sistema, el servidor web manda los datos de la consulta del usuario al *generador de consultas*, mientras que por otro lado recibirá los datos desde el *generador de presentación.*

- **Generador de consultas:** Recibe la consulta realizada por el usuario en forma de cadena de texto de la que forman parte las palabras clave, frases u otros atributos, o bien a través del envío de datos desde un formulario web. Cuando el generador ha recibido estos datos, los interpreta y genera una nueva consulta en formato nativo para que se pueda realizar la búsqueda dentro del repositorio.

- **Buscador:** Se encargará de acceder a la estructura interna de datos (repositorios, índices, etc.) para cumplir el pedido. Un buen buscador debe poder ordenar los resultados de forma que las páginas más relevantes aparezcan primero atendiendo a varios indicadores:

    - **Localización:** Muestra antes los documentos que contienen ocurrencias de todos los términos empleados en la consulta. La relevancia de los documentos será mayor cuanto más al comienzo del mismo se encuentren los distintos términos buscados. Por ejemplo, si un documento contiene en su título todos los términos de la consulta, éste será muy relevante y se mostrará antes en la lista de resultados.

    - **Frecuencia de aparición:** Cuanto más aparezcan los términos de la consulta en un documento, más relevante será éste. Algunos motores controlan el número de ocurrencias permitidas en un documento estableciendo un límite, para así evitar documentos spam que no muestren un valor real de relevancia.

    - **Popularidad:** Consiste en medir la popularidad de la página mediante el número de enlaces que apuntan a ella. Normalmente, una página a la que se realizan muchas referencias suele ser mejor que otra a la que no se hacen tantas.

- **Precio:** Propio de buscadores comerciales. Consiste en implantar servicios de pago que permitan que una página se muestre primero en la lista de resultados en función de la cantidad de dinero pagada.

- **Generador de presentación:** Partiendo de los datos generados por el buscador, construye la vista que se incluirá en el documento HTML mostrado al usuario a través del servidor web.

A continuación se muestra una figura con los componentes definidos anteriormente y cómo se relacionan unos con otros.

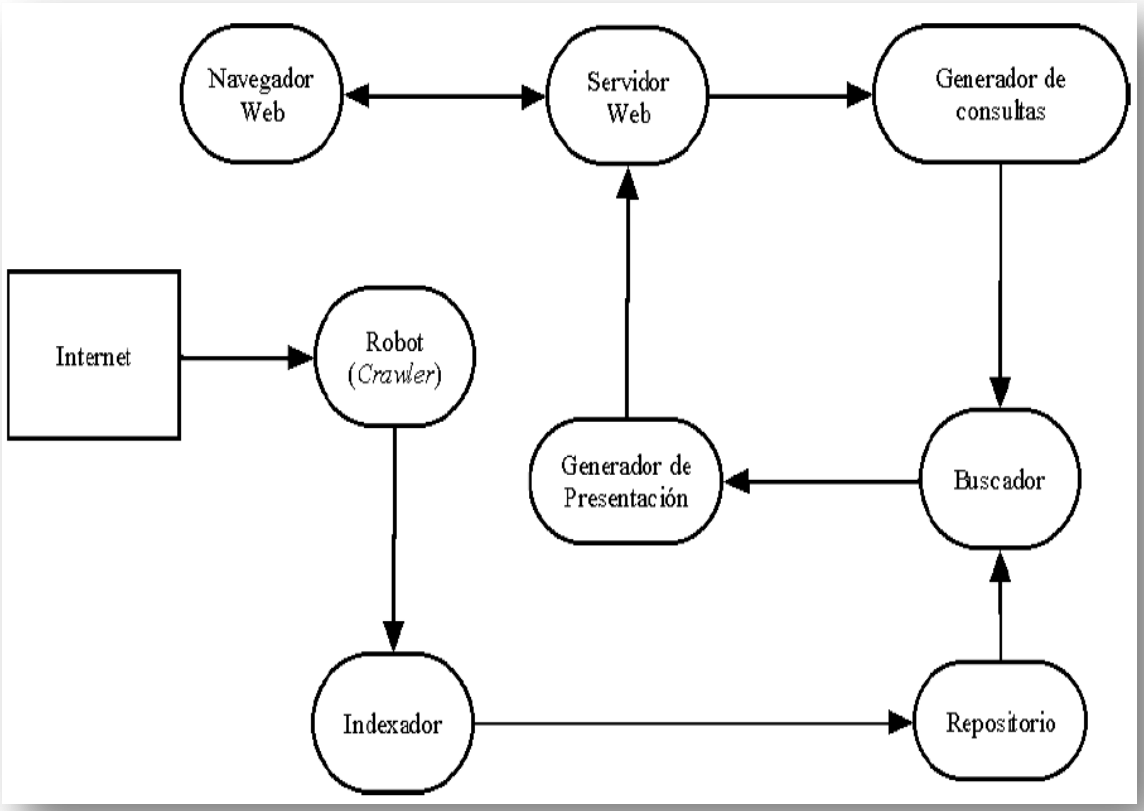

Ilustración 3. Arquitectura de un Motor de Búsqueda

3.5.- Modo de Operación de un Motor de Búsqueda

Las operaciones que puede llevar a cabo un motor de búsqueda se dividen en tres secciones: las operaciones que realiza sobre la consulta, las operaciones que realiza sobre los términos y las operaciones que realiza sobre los documentos.

3.5.1.- Operaciones de consulta

En los motores de búsqueda las consultan se expresan mediante sentencias formales dependiendo de la necesidad de información de los usuarios del sistema.

El *Parsing* es una de las operaciones más comunes y consiste en dividir la consulta en los elementos que la constituyen. Se tienen que dividir las búsquedas booleanas en los términos correspondientes de la indización o palabra clave y los operadores asociados a ellas para formular la expresión de la consulta. Se recupera el conjunto de los documentos asociados a cada término de consulta y estos conjuntos se combinan de acuerdo a los operadores booleanos.

Por otro lado, también encontramos una operación denominada *Reutilización* cuya función es reutilizar una búsqueda efectuada con anterioridad.

3.5.2.- Operaciones sobre los términos

Las operaciones que se llevan a cabo sobre los términos en un motor de búsqueda son las siguientes:

- **Stemming:** Procedimiento para cortar las palabras reduciéndolas a su forma de raíz más común. Existen varios algoritmos de stemming usados en sistemas de recuperación de información. Este procedimiento consigue un aumento del *recall* que se trata de una medida sobre el número de documentos que se pueden encontrar en una consulta.

- **Truncamiento:** Método que consiste en el corte de las palabras pero realizado por el usuario manualmente en los procesos de recuperación de la información.

- **Tesauro:** Esta operación muestra una lista de términos, sus términos sinónimos y las relaciones semánticas mantenidas entre los términos del mismo.

- **Palabras vacías (o stop):** Como se ha mencionado anteriormente, estas palabras no aportan ningún significado a la búsqueda y por lo tanto no hay que tenerlas en cuenta.

- **Ponderación de términos:** Proceso que consiste en asignar un valor numérico basado en su distribución estadística, es decir, en la frecuencia con la que los términos aparecen en documentos, colecciones de documentos, etc.

- **Lematización:** aplica heurísticas para eliminar flexiones y afijos, utilizando el análisis morfológico para determinar el tipo de palabra y aplicar reglas en consecuencia para obtener la base de la palabra (lema).

3.5.3.- Operaciones sobre documentos

Los documentos son los principales objetos de un buscador y, por lo tanto, existen muchas operaciones para ellos. La operación común consiste en ordenar los documentos recuperados tras las consulta por algún campo determinado.

3.6.- Algoritmos usados en los Motores de Búsqueda

Podemos agrupar en cuatro topologías fundamentales los algoritmos usados por los buscadores para mostrar las páginas de resultados.

- **Modelo booleano:** Este modelo está basado en la lógica de proposiciones de George Boole [3]. Plantea la presencia o ausencia de términos sin tener en cuenta el contexto. Las relaciones entre conceptos se pueden expresar como relaciones entre conjuntos y, de esta forma, las ecuaciones de búsqueda pueden transformarse en ecuaciones matemáticas que ejecutan operaciones sobre estos conjuntos, dando como resultado un nuevo conjunto. Las consultas se llevan a cabo mediante operadores booleanos (AND, OR, NOT). La combinación de estos operadores permite realizar búsquedas complejas, aunque la mayoría de los usuarios suelen tener problemas con estos operadores.

- **Modelo de espacio vectorial:** Es un modelo algebraico desarrollado por Gerald Salton que se usa para filtrado, recuperación, indexado y cálculo de relevancia de información. Se basa en la frecuencia de aparición de los términos. Este modelo consiste en que las distancias y direcciones entre palabras y frases extraídas del texto se miden en un espacio multidimensional, representando cada documento o consulta como un vector en un espacio n-vectorial. El número de términos únicos presentes en el cuerpo del documento, son los que determinan esta dimensión. Las palabras significativas se eliminan del vector y se incluyen en un listado de palabras vacías con el objetivo de reducir el porcentaje de palabras con mayor frecuencia de aparición. A continuación, se asignan pesos a los términos con el fin de indicar el grado de importancia en la representatividad del documento. Para finalizar, se aplica el coeficiente de similitud, lo que quiere decir que los vectores de dos documentos estarán más cerca si tienen más términos en común.

- **Modelo probabilístico:** Se basa en el cálculo de la probabilidad de que un documento se corresponda con una pregunta. Este modelo sigue las bases del modelo probabilístico establecido por Robertson y Sparck Jones en [4]. Su objetivo es calcular la probabilidad de que un documento sea relevante para la consulta dado que dicho documento posea unas determinadas características, características en forma de términos índice que contiene dicho documento. Según el principio de orden de probabilidades [5], el rendimiento óptimo de un sistema se consigue cuando los documentos se ordenan de acuerdo a sus probabilidades de relevancia.

- **Modelo hipertextual basado en la conectividad de enlaces:** Es un modelo cuya función es tener en cuenta la propia estructura hipertextual basándose no sólo en el recuento de enlaces, sino también en el análisis de éstos y las relaciones que establecen. Asimismo, cada vez es más común tener en cuenta la calidad y origen de estos enlaces.

Los algoritmos definidos anteriormente no son excluyentes entre sí y, por lo tanto, muchos motores de búsqueda combinan varios modelos.

3.7.- Importancia de los Motores de Búsqueda

Cada día, millones de personas en todo el planeta encienden el ordenador o utilizan el teléfono móvil desde cualquier lugar para buscar información en la red. ¿Qué hacen para buscar esta información? Se meten directamente en la página de algún motor de búsqueda, no importa el que sea.

Se estima que sólo en inglés, considerado el idioma predominante en la red, se hacen diariamente más de 500 millones de búsquedas, por lo que hay que imaginar cuál será el número total de búsquedas en todo el planeta. Además, el crecimiento anual de estas búsquedas a nivel mundial es del 39%. De estas búsquedas, la mayoría de ellas (casi el 90%) se realizan únicamente en tres motores de búsqueda: Google, Yahoo y MSN (Bing) respectivamente.

Por otro lado, cada vez son más exigentes los usuarios de Internet a la hora de buscar la información que les interesa, ya que, si no la encuentran en las dos o tres primeras páginas de los resultados del motor, modifican los términos de la consulta o incluso cambian de buscador, rechazando el resto de páginas web que se encontraban en la lista de resultados en páginas siguientes. Además, los motores de búsqueda varían sus algoritmos de búsqueda frecuentemente para brindar información más relevante a los usuarios. Otro factor importante es el continuo crecimiento de páginas web que se suman a las ya existentes en Internet, lo que hace que cada vez haya un mayor número de páginas entre las que seleccionar los mejores resultados.

Los buscadores son las funciones más utilizadas por los usuarios de internet después del correo electrónico, tal y como se puede ver en la siguiente gráfica.

	18-33 Gen. Milenio	34-45 Gen. X	46-55 Gen. Jov Boom	56-64 Gen. May. Boom	65-73 Gen. Silen	74+ Gen. G.I	
1	Correo-e	Correo-e	Correo-e	Correo-e	Correo-e	Correo-e	
2	Búsquedas	Búsquedas	Búsquedas	Búsquedas	Búsquedas	Búsquedas	
3	Info salud	Info salud	Info salud	Info salud	Info salud	Info salud	
4	Redes Sociales	Leer noticias	Leer noticias	Leer noticias	Leer noticias	Comprar	
5	Ver vídeos	Webs gob.	Webs gob.	Webs gob.	Reserv viajes	Leer noticias	
6	Leer noticias	Reserv viajes	Reserv viajes	Comprar	Comprar	Reserv viajes	
7	Comprar	Ver vídeos	Comprar	Reserv viajes	Webs gob.		90-100%
8	Mensaj. Inst	Comprar	Ver vídeos	Banca online			80-89%
9	Esc. Música	Redes Sociales	Banca online	Ver vídeos			70-79%
10	Reserv viajes	Banca online	Redes Sociales				60-69%
11	Anuncios	Anuncios					50-59%
12	Banca online	Esc. Música					
13	Webs gob.	Mensaj. Inst					

Figura 7. Usos de Internet

Fuente: http://www.ub.edu/blokdebid/es/content/informe-pew-sobre-el-uso-de-internet-por-generaciones-2010

Dentro de las búsquedas on-line, el 50% abordan temas relacionados con los negocios y el 41% de las compras generadas por Internet han surgido desde los buscadores. Por otra parte, las compras off-line también están influenciadas por los motores de búsqueda, tanto que 3 de cada 4 usuarios europeos afirman que antes de hacer una compra off-line han buscado información sobre productos y servicios on-line.

Esta información nos demuestra que millones de usuarios buscan frecuentemente productos y empresas a través de los buscadores, por lo que las empresas tienen el objetivo de conseguir un buen posicionamiento en los resultados de las búsquedas para conseguir aumentar de esta forma su número de clientes, ya que, como he mencionado anteriormente, la mayoría de los usuarios sólo miran las tres primeras páginas de los resultados de búsqueda.

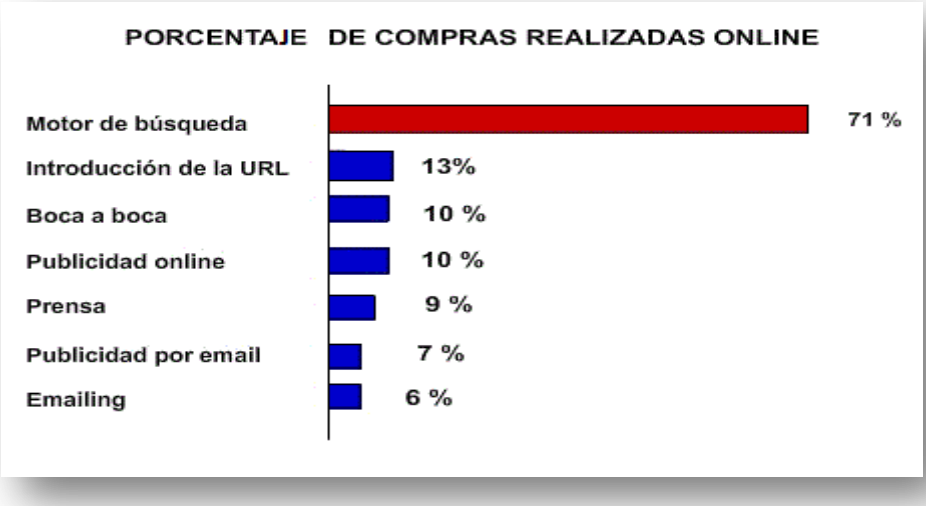

Figura 8. Porcentaje de compras realizadas online (Fuente: Double clic)

Debido a la importancia del buen posicionamiento web en los resultados de búsqueda, las empresas utilizan varias técnicas para conseguirlo, ya pueda ser pagando una cantidad de dinero a los buscadores o utilizando diferentes herramientas SEO (Search Engine Optimization).

Algunos aspectos que muestran la importancia de conseguir un buen posicionamiento en los motores de búsqueda son los siguientes:

- El 90% de los usuarios que buscan información o recursos en Internet, utilizan los motores de búsqueda para dicho fin.
- El 84% de los usuarios que usan Google.com sólo prestan atención a la primera página de resultados.
- El 65% de los usuarios que usan Google.com nunca han hecho clic en los enlaces patrocinados (parte de pago).

3.8.- Posicionamiento en Buscadores (SEO)

Una de las finalidades principales de la mayoría de sitios webs, es aparecer lo más arriba posible en la lista de resultados de los motores de búsqueda para así aumentar su número de visitantes y futuros clientes. Para ello, muchas páginas utilizan herramientas SEO.

SEO es el proceso de mejorar la visibilidad de una página web en los motores de búsqueda sin pagarles dinero para conseguir acceder a una posición destacada en la lista de resultados. Dicho posicionamiento se consigue realizando tareas de optimización en las páginas web con el objetivo de aumentar el número de visitas en las mismas para así aparecer en los primeros resultados de búsquedas.

Ya que el uso de estas técnicas puede afectar a los resultados naturales de las búsquedas, los motores de búsqueda importantes regulan las técnicas SEO aplicadas por cada página web para evitar que incumplan una serie de cláusulas y condiciones de uso establecidas por los mismos que pueden considerar dichas páginas como una forma de spam (spamdexing).

El término SEO involucra tanto al código de programación y diseño como a los contenidos y personas encargadas de realizar este trabajo.

3.8.1.- Importancia de un buen posicionamiento. Estudios de Eye-Tracking.

Uno de los temas más discutidos en la comunidad SEO es la forma en la que los usuarios interactúan con la página de resultados, y es también un campo de estudio muy importante por los especialistas en el posicionamiento Web. Para conocer esta interacción se realizan experimentos de “Eye-Tracking”.

El objetivo del los estudios “Eye-Tracking” es obtener información sobre cómo los usuarios navegan a través de unos patrones abstractos y realizan clic en los enlaces. Estos estudios usan cámaras especiales parecidas a los lectores de códigos de barras, que graban los movimientos de nuestros ojos y atienden a una serie de indicadores establecidos previamente. Acciones como por ejemplo cómo se mueven nuestros ojos, dónde se paran, cuándo aumenta la pupila y otros muchos parámetros, nos permiten entender cómo navegamos tanto consciente como inconscientemente.

El primer estudio publicado que utiliza la técnica de “Eye-Tracking” para analizar el comportamiento de los usuarios en las SERPs es el realizado en la Universidad de Cornell (USA) por Laura A. Granka, Thorsten Joachims y Geri Cay [6], aunque en este estudio no se tiene en cuenta la intención de búsqueda del usuario. Dicho estudio proporcionó los siguientes resultados:

Ilustración 4. % Distribución de clics en SERPs.

En la imagen anterior podemos que ver que el primer resultado es el que se lleva la mayor dedicación de tiempo y el mayor número de clics. Con respecto al segundo resultado de búsqueda, presenta un brusco descenso del número de clics respecto al primer resultado, aunque el porcentaje de tiempo dedicado es más o menos similar. Después el porcentaje sigue disminuyendo a pesar de algunas pequeñas subidas.

Otro dato interesante que podemos sacar de la imagen es que la posición 7 en la lista de resultados, en la que menos clics obtiene de todas, menos aún que las posiciones posteriores. Esto se debe a que la posición 7 es frecuentemente ignorado por el scroll en el movimiento de la página que realiza el usuario.

Otro estudio realizado por la empresa Eyetools en 2005 demostró que el mayor número de usuarios dirigían su vista a un triángulo ficticio, denominado triángulo de oro, que cubre los primeros resultados de búsqueda, tal y como se muestra en la siguiente imagen:

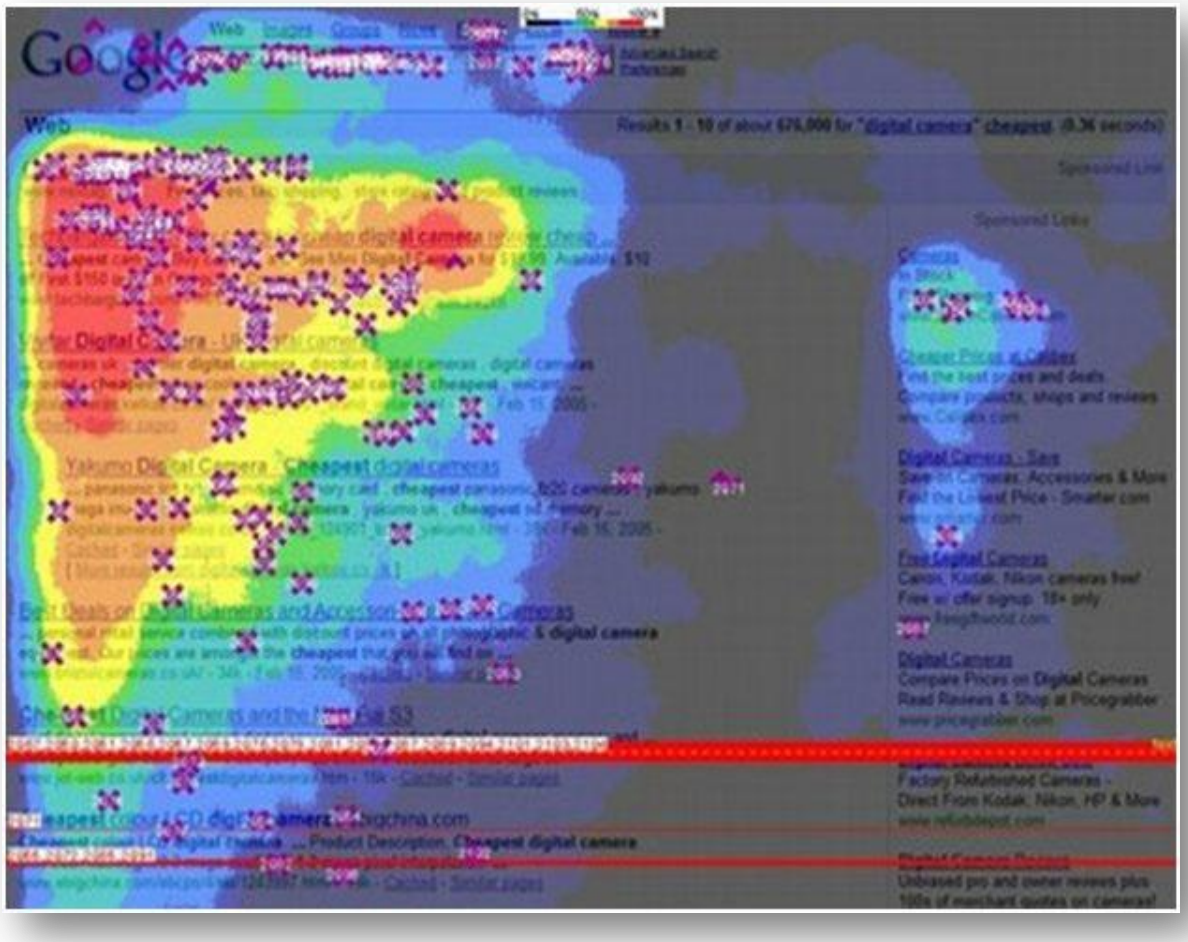

Ilustración 5. Triángulo de Oro de las SERPs.

En la imagen se puede ver que cuanto más arriba aparece el resultado, mayor es el porcentaje de usuarios del buscador que lo ven.

En 2010, un estudio realizado por Mari-Carmen Marcos y Cristina González-Caro en la Universidad Pompeu Fabra de Barcelona viene a romper ese ‘triángulo de oro’ de las búsquedas [7]. Lo que busca este estudio, es hacer una valoración de cuáles son los hábitos visuales de los usuarios a la hora de utilizar los resultados de búsqueda obtenidos a partir de los buscadores Google y Yahoo!, atendiendo al tipo de consultas que realicen. Estas consultas pueden ser:

- **Informacional:** el usuario pretende obtener una información determinada. En este caso, el usuario se centrará en los snippets para determinar si la página web puede mostrarle la información que busca.

- **Navegacional:** el usuario quiere llegar a un sitio web determinado. Se centrará en el título de la página para ver si coincide con el del sitio que

busque. En caso de ser un usuario avanzado, también se fijará en la URL del sitio para ver si se trata de la página oficial.

- **Transaccional:** el usuario quiere realizar una acción. En este caso prestará más atención a los enlaces patrocinados (título del anuncio) y en los resultados orgánicos, mirará tanto el título como el snippet.

- **Multimedia:** son aquellas búsquedas que tienen como finalidad ver alguna foto o vídeo. El usuario centrará prácticamente toda la atención en la imagen asociada e ignorará el resto de información.

Los resultados obtenidos tras el estudio atendiendo al tipo de búsqueda fueron los siguientes:

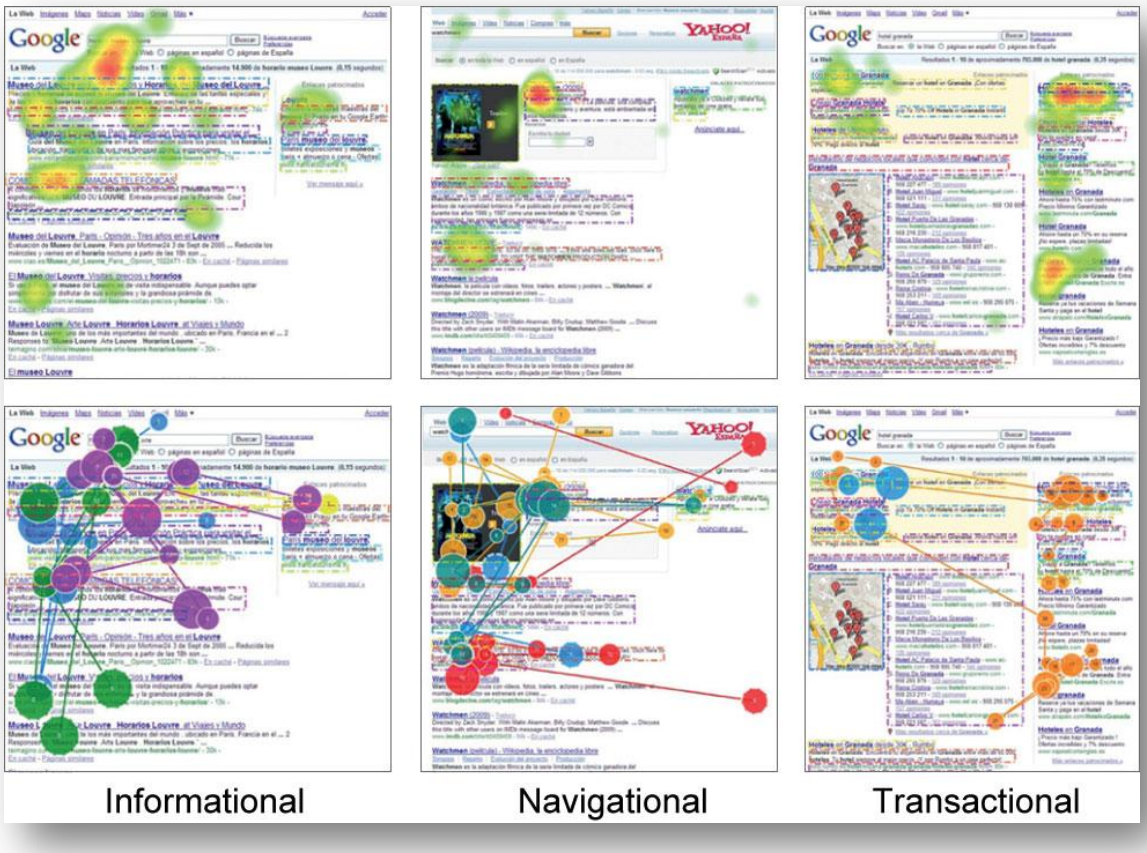

Ilustración 6. Comportamiento según tipo de búsqueda

Algunas de las conclusiones sacadas de este estudio son que los resultados orgánicos siguen recibiendo más atención que los patrocinados (82% frente al 17%), dentro de éstos se conserva el orden de atención en las áreas de interés: snippet, título y url.

Cuando el objetivo es transaccional, los enlaces patrocinados atraen más la atención que en el caso de las búsquedas informacionales o navegacionales, aun así, la atención que reciben es bastante inferior a la de los resultados orgánicos.

Por otra parte, las consultas multimedia reciben la mayor atención en sus imágenes:

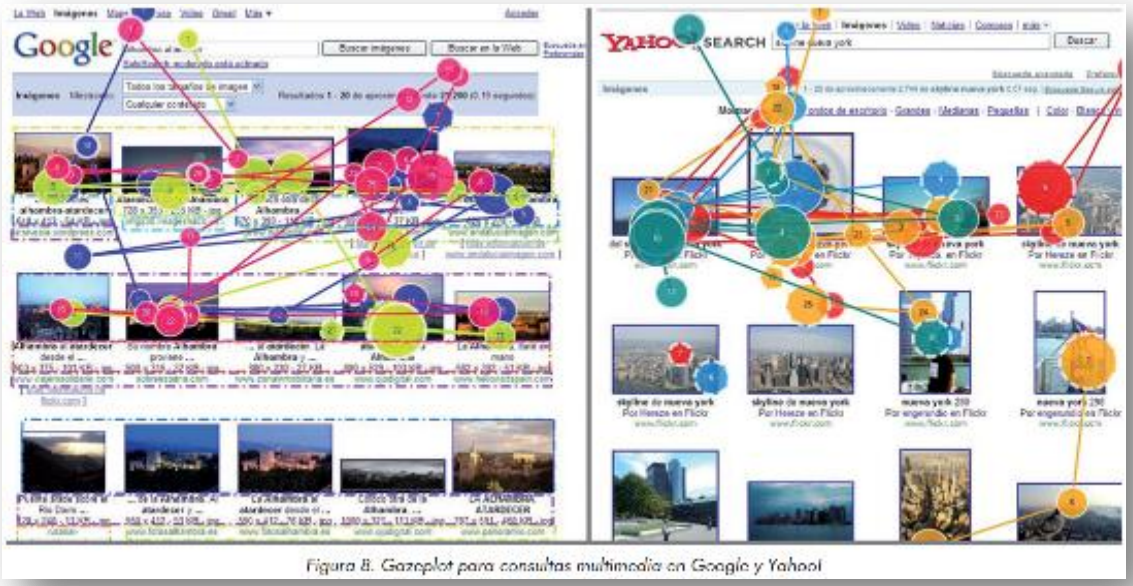

Ilustración 7. Comportamiento en consultas multimedia.

Otras conclusiones extraídas de este estudio son que los anuncios situados en la zona superior reciben más atención que los que se encuentran en la zona lateral, y que el snippet es el área de resultados que más atención reciben por parte de los usuarios.

3.8.2.- Cómo mejorar el posicionamiento

Algunas de las acciones que se pueden realizar para conseguir un mejor posicionamiento en las páginas de resultados son las siguientes:

- Lograr que otras páginas de contenido o temática similar enlacen con tu sitio web.
- Registrarse en directorios importantes que, aunque hayan perdido interés, son buenos para conseguir enlaces o un primer rastreo de tu sitio web por los motores de búsqueda.
- Participar en foros, principalmente en aquéllos que estén relacionados con la temática de tu sitio web. La participación debe aportar contenido valioso.
- Hacer aportaciones en otros sitios web, ya que es un buen método para conseguir atraer un mayor número de visitas.
- Participar en redes sociales para darse a conocer y obtener nuevas visitas.

- Incluir en los contenidos de tu sitio web alta densidad de palabras clave que usan frecuentemente los usuarios y potenciales clientes para realizar las búsquedas.

- Poner las palabras clave que se pretenden posicionar en negrita o cursiva.

- Usar las etiquetas “meta” con las palabras clave.

- Para destacar títulos o términos importantes hacer uso de las cabeceras h1, h2, etc. Usar palabras clave en las cabeceras.

- Hacer que el diseño de nuestra página web sea funcional, fácil de acceder y que capte la atención del usuario para mejorar el posicionamiento.

- Limitar el contenido en Flash, frames o JavaScript ya que no permiten el rastreo por parte de los robots.

- Enlazar internamente las páginas de nuestro sitio de forma ordenada. Incluir en el código un “mapa del sitio” dará paso al buscador por las diferentes secciones de la página mejorando su visibilidad.

- Actualizar frecuentemente el contenido del sitio web con información de calidad.

3.8.3.- Nuevas tecnologías de SEO

Recientemente están surgiendo en los buscadores nuevas tecnologías de SEO que han incorporado nuevas variables a tener en cuenta para la optimización de un sitio web. Algunas de estas nuevas tecnologías son:

3.8.3.1.- Búsquedas Universales

El uso de estas técnicas es uno de los cambios más notables en los motores de búsqueda, ya que en las páginas de resultados se muestran de forma combinada resultados orgánicos con imágenes, videos, blogs, etc. Por esta razón, hay que adaptar la optimización para todas estas búsquedas universales. Esto tiene vital importancia para el contenido vertical de una página en un tema concreto.

3.8.3.2.- Búsquedas Personalizadas

Las búsquedas personalizadas son aquellas que los usuarios realizan de forma más frecuente y eligen sus páginas favoritas. Google ha lanzado algunas características que facilitan a los buscadores enlazar contenido con las acciones del usuario en el buscador que se encuentran almacenadas en el Google Webhistory, el

Google Bookmarks y Google Sidewiki. Además, ahora aparecen estrellas en el buscador que permiten que los usuarios marquen los enlaces como favoritos.

Esta búsqueda personalizada de Google sólo era para los usuarios registrados que tenían habilitado el historial Web, sin embargo, ahora todos los usuarios, tengan o no cuenta de Google, tienen resultados de búsqueda personalizados. El método que emplea Google para realizar esta acción es mediante cookies anónimas que se guardan en el disco duro de los usuarios y que almacenan la información de su actividad en los últimos 180 días. Este historial, está desvinculado de la cuenta de Google y su historial Web.

3.8.3.3.- Búsquedas en Tiempo Real

Otorga la facilidad de enlazar los resultados de una búsqueda con sitios web que proveen contenido en tiempo real, como lo son las redes sociales Twitter, MySpace o Facebook. Se prevé que aparte de las mencionadas anteriormente, también se incluyan Google Buzz y Google Wave cuando sea lanzado oficialmente.

3.8.3.4.- Búsquedas Sociales

A finales del 2009, Google lanzó como experimento las búsquedas sociales, cuyo propósito es suministrar contenido proveniente de las redes sociales escritos en el perfil de usuario en las páginas de búsqueda. Estos contenidos se obtienen a partir de las palabras clave que un usuario tiene en la red social y además son relevantes con la búsqueda realizada.

3.8.3.5.- Búsquedas Locales

Es una característica que presenta Google que consiste en una combinación de Google Maps con Google Local Business y proporciona un determinado contenido atendiendo al lugar en el que estemos realizando la búsqueda. En ella los resultados obtenidos tienen validez en una región geográfica determinada. Se puede optimizar enviando los datos de un negocio local en el Google Local Business Center.

3.9.- Limitaciones de los Buscadores

El estudio realizado por Michael Bergman para Bright Planet [8] estima que el 85% de los usuarios de Internet, utilizan los motores de búsqueda para satisfacer sus necesidades de información. Sin embargo, a pesar de la gran utilidad de los buscadores, también es conocido que dichas herramientas tienen una serie de limitaciones, lo que lleva a que un gran número de usuarios no encuentre lo que necesita cuando los utilizan.

Tal como afirma Isidro Aguillo [9], la incapacidad de los motores para cubrir la inmensa cantidad de información web disponible en la red y que aumenta cada día, es una de las principales limitaciones que presentan los buscadores. Isidro Aguillo, además, afirma que los grandes motores de búsqueda (como Google o Altavista) apenas abarcan el 20-25% de los contenidos presentes en la red.

Para Silvia Osorio [10], una de las gerentes del portal Internet Invisible, los buscadores convencionales únicamente pueden buscar en un 10% de la red. Por esta razón, cuando los usuarios utilizan los buscadores para localizar información en Internet, no están buscando en toda la red, sino en la base de datos del buscador que estén utilizando, por lo que están “ignorando” una gran fuente de información.

Otro de los problemas que presenta el uso de los buscadores, según Enrique Alfonso [11], es que muchos de ellos únicamente pueden localizar e indizar páginas html, por lo que los archivos de imágenes, audio, video, pdf y otros formatos no pueden ser localizados por ellos. La razón por la que muchos buscadores no indexan estos formatos es porque hay menos demanda de ellos y, además, son más difíciles de archivar y organizar, requiriendo mayores recursos del servidor [12]. Este problema no está presente en todos los buscadores, algunos como Google o Altavista ya permiten recuperar este tipo de formatos, aunque aún siguen siendo incapaces de recuperar toda la información almacenada en la red, como la almacenada en servidores que no permiten el acceso público.

Por otro lado, de la mano de la Web 3.0, también conocida como Web Semántica, llega una nueva tendencia que busca la renovación de todos los buscadores para incorporar la búsqueda semántica. Este concepto busca dar más valor al contexto y significado de las palabras dejando más de lado los conceptos de meta etiquetas, keywords y demás. Muchos investigadores hablan de este tipo de búsqueda, como una búsqueda más “inteligente”, en la que se busca que los motores de búsqueda sean menos “robóticos” y se conviertan en máquinas que puedan interpretar de la forma más efectiva posible lo que busca el usuario y, así, poder mostrar los resultados más apropiados.

La gran ventaja que presentan los buscadores semánticos es que, además de tener la capacidad de interpretar mejor los textos de los sitios web, también pueden realizar una interpretación más precisa de lo que el usuario está buscando información.

4.- Archivos Robots.txt

4.1.- Definición

Un archivo robots.txt es un archivo de texto sencillo en el cual se establecen las recomendaciones de indexación de un sitio para las diferentes arañas de búsqueda (spiders) que recorren la web. Cuando una araña llega a un sitio web, su primera acción es buscar el archivo robots.txt en la raíz de la web para conocer qué páginas puede indexar y si hay alguna carpeta o directorio que no debería recorrer. Por ejemplo, esto se puede usar para dejar fuera de una preferencia los resultados de una búsqueda avanzada, o de la creencia que el contenido de los directorios seleccionados puede ser engañoso o inaplicable a la clasificación del sitio en su totalidad.

Los servicios que este fichero permite y los beneficios que aporta son los nombrados a continuación:

- **Impedir el acceso a determinados robots:** En algunas ocasiones puede interesarnos impedir el acceso a nuestra página a determinados robots de búsqueda, por ejemplo en el caso de estar construyendo una página web nueva y no queremos que sea visible. Además, hay algunos robots que no pertenecen a buscadores y pueden causarnos algunos problemas (robots que pretenden hacer spam).

- **Prohibir el acceso a determinadas zonas de la web:** Puede darse el caso de que nos interese tener alguna zona de la web que no queramos que sea indexada por los buscadores. De esta manera se puede evitar, por ejemplo, la recuperación de contenidos superfluos, como ficheros de logs (ficheros encargados de registrar las entradas al sitio), el acceso a contenidos duplicados, o impedir la recuperación de contenido que no queremos que aparezca en el buscador. Hay que tener claro que esto no prohíbe el acceso a ciertas zonas de la página a los usuarios de la misma, sino que evita que aparezca en los buscadores.

- **Reducir la sobrecarga del servidor:** Se podrá controlar el flujo de información de algunos robots, ya que varios de ellos realizan un número elevado de peticiones que pueden llegar a saturar tu servidor.

- **Establecer la frecuencia de paso de los robots:** Este punto está relacionado con el anterior, ya que, si podemos detectar que algún robot envía demasiadas solicitudes a nuestro sitio web, podemos ver que puede llegar a saturar el servidor.

- **Eliminar contenido duplicado:** La importancia de esta acción radica en que si conseguimos eliminar la duplicidad de contenido, los buscadores nos puntuarán muy alto y, por lo tanto, aumentaremos el flujo de visitas a nuestra web.

- **Indicar la ubicación del *sitemap.xml*:** En el fichero robots.txt se puede indicar en qué parte se encuentra el mapa del sitio (sitemap), sistema que permite indicar a los buscadores las URLs de todas las páginas de un sitio web.

Al utilizar el archivo robots.txt hay que tener en cuenta dos consideraciones importantes:

- Hay robots de búsqueda que pueden ignorar el archivo robots.txt de un sitio web. Normalmente suele ocurrir cuando se trata de robots de malware, cuyo objetivo es buscar las vulnerabilidades de seguridad de las páginas web, o en el caso de los spammers que buscan recopilar direcciones de correo electrónico.

- El robots.txt es un archivo público disponible. Cualquiera puede ver qué zonas de un sitio web no quiere que indexen los robots de búsqueda su propietario.

Como se ha mencionado en el primer punto, hay algunos robots de búsqueda que pueden ignorar el archivo robots.txt de una página, lo que nos puede llevar a preguntarnos lo siguiente:

¿Hay alguna forma de bloquear el acceso de los robots “malos”?

En teoría, tras lo explicado anteriormente si debería ser posible, pero en la práctica no lo es. En el caso de conocer el nombre del robot “malo” y, teniendo la suerte de que éste obedezca al fichero robots.txt, bastaría con incluirlo en la directiva User-Agent del fichero. Sin embargo, como la mayor parte de los robots “malos” ignoran el robots.txt, aun conociendo el nombre de ese robot, no sería posible impedir su acceso al sitio web.

Si dicho robot operase desde una única dirección IP, podríamos prohibir su acceso al servidor usando determinadas técnicas como los ficheros de iptables o mediante la configuración del servidor.

Por el contrario, si copias del mismo robot operasen desde distintas direcciones IP, la mejor opción sería utilizar una configuración avanzada de reglas de firewall que bloquee el acceso a las direcciones que realicen múltiples conexiones a nuestro servidor, aunque esta técnica podría afectar tanto a robots “malos” como a “buenos”.

4.2.- Estándar de Exclusión de Robots

El Estándar de Exclusión de Robots o en inglés A Standard for Robots Exclusion (SRE), es un método para impedir que los robots de los motores de búsqueda que analizan el contenido de parte o la totalidad de los sitios web,

públicos o privados, incorporen información innecesaria a los resultados de búsqueda. Los motores de búsqueda usan frecuentemente los robots para categorizar archivos de los sitios web, así como los webmasters los utilizan para corregir o filtrar el código fuente.

Este estándar en realidad no es un estándar oficial, sino un acuerdo entre los integrantes de varias listas de correo sobre robots y diversos temas relacionados con la tecnología WWW. Fue creado en junio de 1994 por los miembros que enviaban la lista *robots-request@nexor.co.uk*.

Su misión básica es suministrar un medio para que los webmasters indiquen qué zonas de su sitio web no quieren que sean analizadas, así como qué robots tienen permitido el acceso y cuáles lo tienen prohibido.

Algunas de las desventajas que presenta el estándar SRE es que se trata de un protocolo consultivo. Esto significa que aunque los robots deben estar programados para buscar el fichero robots.txt, leerlo y seguir sus reglas, como este estándar es un simple acuerdo, pueden ignorarlo.

Los webmasters en ocasiones utilizan este fichero para hacer privadas o invisibles algunas partes de su sitio web pero, ya que este fichero es público, el contenido del mismo puede ser visto por cualquier usuario de la web.

Este estándar no incluye las directivas **Allow**, **Crawl-delay** y *, aunque están incluidas en muchos ficheros robots.txt. Por esta razón, hay robots que reconocen e identifican dichas directivas mientras que otros no las tienen en cuenta.

Es importante que todos los robots intenten respetar este estándar ya que, de no hacerlo, podrían darse situaciones comprometidas, como por ejemplo, que un robot envíe repetidas peticiones sobre un determinado recurso llegando a causar la caída del servidor por sobrecarga, o incluso saturar un router. Esto puede ser un motivo para que los webmasters incluyan ese determinado robot en la lista de *visitantes no deseados*.

El estándar puede implementarse analizando los robots.txt situados en la raíz del dominio del servidor. También hay *meta-tags* que pueden incluirse dentro del código fuente de las páginas web.

4.3.- Como crear un archivo robot.txt

4.3.1.- ¿Cómo se crea?

La creación de un archivo robots.txt es bastante sencilla, simplemente basta con abrir un editor de texto y definir las directivas que se mostrarán más adelante. También existen diferentes herramientas que nos permiten generar estos ficheros de forma automática como la ofrecida por Mcanerin (http://www.mcanerin.com/EN/search-engine/robots-txt.asp), la herramienta

webmaster de Google (www.google.com/webmasters/) o la perteneciente a Seobook (http://tools.seobook.com/robots-txt/generator/).

4.3.2.- ¿Dónde ponerlo?

El archivo robots.txt tiene que estar colocado en el directorio raíz de la página web. Esto se debe a que cuando un robot realiza una búsqueda del fichero robots.txt en la URL de una página, lo que hace es borrar todo lo que viene después de la primera barra (/) y poner /robots.txt. Por ejemplo si la URL de mi página es www.mipagina.es/carp/index.html, el robot eliminará "/carp/index.html" y lo reemplazará por "/robots.txt" quedando de la siguiente forma www.mipagina.es/robots.txt.

Un aspecto importante a recordar es que el nombre del fichero robots.txt debe ir siempre en minúsculas.

4.3.3.- ¿Qué hay que poner?

El archivo robots.txt se compone de dos directivas principales:

- **User-agent:** En esta directiva debemos incluir el nombre o nombres de los robots a los que queremos impedir la indexación de partes o la totalidad de nuestro sitio web. En el caso de que queramos impedir la indexación a todos los robots, podemos usar el asterisco *.

- **Disallow:** Determina el acceso a los distintos sitios de la página. Hay tres formas principales:

    - **Disallow:** permite el acceso a todo el sitio web.
    - **Disallow: /** prohíbe la entrada a todo el sitio web.
    - **Disallow: /carpeta/** prohíbe la entrada a todos los documentos que se encuentren dentro del directorio carpeta.

En algunos casos se utiliza la palabra **Allow** en lugar de **Disallow** para indicar qué accesos están permitidos, aunque es mejor no utilizarla ya que algunos crawlers no entienden la palabra **Allow**. Además, se entiende que las rutas omitidas en la directiva **Disallow** están permitidas por defecto.

Por otra parte, está permitido utilizar varias directivas **Disallow** bajo el mismo **User-agent**, mientras que no se pueden usar varios **User-agent** bajo un mismo **Disallow**.

Añadiendo al principio de una línea el carácter almohadilla # podemos escribir comentarios que el crawler no interpretará.

Esta permitido ir acumulando reglas para diferentes crawlers, por lo que cada vez iremos formando un archivo robots.txt más largo y completo. Se debe

dejar una línea en blanco de separación cada vez que escribamos una directiva **User-agent**. Asimismo, existe una ligera adaptación en algunos crawlers que nos permites usar operadores de truncamiento (? y *) en las rutas para hacer más exhaustivo el comando. También se puede usar el signo $ para indicar que el texto seleccionado tiene que aparecer al final de la URL. Esto se usa normalmente para indicar las extensiones que no se quiere que sean indexadas, por ejemplo **Disallow:/*.pdf$** hará que los robots no indexen ningún fichero con extensión .pdf.

Otro de los aspectos a tener en cuenta es a la hora de usar cosas como **Disallow:/contenido** o **Disallow:/*contenido,** ya que en vez de impedir el paso al directorio ‘contenido’ como queríamos, estamos impidiendo el acceso a direcciones como **/mostrar-mi-contenido** o **/contenido-de-la-exposicion**.

4.3.4.- Ejemplos de robots.txt

A continuación mostraremos una serie de ejemplos de robots.txt dependiendo del tipo de funcionalidad que queramos otorgarle.

Permitir el acceso a toda la página a todos los buscadores

User-agent: *
Disallow:

Impedir el acceso a toda la página a todos los buscadores

User-agent: *
Disallow: /

Impedir el acceso a toda la página a un buscador determinado

User-agent: Googlebot
Disallow: /

Impedir el acceso a varios directorios de la página

User-agent: *
Disallow: /imágenes/
Disallow: /videos/
Disallow: /articulos/

Impedir el acceso a ficheros determinados de la página a cualquier buscador

User-agent: *
Disallow: /artículos/universidad.pdf
Disallow: /imágenes/escudo.jpg

Impedir el acceso a ciertas carpetas o a toda la web a ciertos robots:

User-agent: *
Disallow:

User-agent: Googlebot-Mobile
Disallow: /imagenes
Disallow: /logs

User-agent: Googlebot-Image
Disallow: /*.php$

Las restricciones hay que hacerlas o con todos los robots o con cada uno individualmente, pero no se pueden incluir varios robots en la misma línea de **User-agent.**

Como se ha explicado anteriormente, la última línea indica que el robot Googlebot-Image tiene el acceso restringido a cualquier fichero con extensión .php.

Permitir el acceso a toda la página a un único buscador

User-agent: msnbot
Disallow:

User-agent: *
Disallow: /

Establecer la frecuencia de paso a un buscador (en segundos)

User-agent: Slurp
Crawl-delay: 30

Con esta directiva le estamos indicando al robot de Yahoo que tiene que esperar 30 segundos entre cada acceso. Esta es una manera de evitar que determinados robots fundan a peticiones a nuestro servidor y puedan llegar a saturarlo.

Incluir un mapa del sitio

Sitemap: http://www.mipagina.com/sitemap.xml

Protocolo de exclusión por medio de meta etiqueta ROBOT

Puesto que los archivos robots.txt sólo podemos usarlos cuando tenemos acceso al servidor, existe otra forma de conseguir la misma funcionalidad utilizando meta etiquetas robots. La meta etiqueta robot permite al creador o al encargado de posicionamiento en motores de búsqueda indicar al crawler que no queremos que indexe una página determinada y, de igual forma, indicarle si queremos que siga los vínculos presentes en la misma o no.

Esta meta etiqueta la tenemos que colocar únicamente en las páginas que no queremos que sean indexadas, en el resto no es necesario ni recomendable. La sintaxis de la meta etiqueta robots es la siguiente:

<meta name=“robots” content=“robots-terms”>

El lugar de colocación de la misma será en la parte head de nuestro documento, junto con las meta **description** y **keywords**. Como se puede observar, la meta etiqueta está formada por dos instrucciones, por un lado tenemos **name,** que nos indica que se trata de una meta robots para indicar la indexabilidad o no de la página (en la instrucción **name** también podemos indicar el nombre de algún robot en concreto), y por otro lado tenemos **content,** que señala la acción que queremos que se desempeñe.

El contenido dentro de robots-terms es una lista con uno o varios de los siguientes indicadores separados por comas:

- **Noindex / Index:** indica al crawler si esta página puede ser indexada o no. A diferencia de la directiva **Disallow** de robots.txt, si elegimos el operador **noindex** conseguiremos que la página no aparezca de ninguna manera en los resultados de búsqueda.

- **Nofollow / Follow:** este término le indica al crawler si tiene permitido recorrer o seguir recorriendo la web a través de los enlaces que encuentre en el cuerpo del documento.

- **Noarchive / Archive:** este operador nos permite decir si queremos que el motor de búsqueda archive o no el contenido de la página en su caché interna. Según Google, este indicador no evita que se guarde en caché la página, sino que no permite que los usuarios del buscador la vean, por lo tanto no se muestra el enlace en la lista de resultados.

- **Nosnippet / Snippet:** este indicador sirve para que el motor de búsqueda no muestre ninguna descripción de un sitio web, simplemente que muestre el título. Si usas **nosnippet** se define automáticamente un **noarchive** y, por lo tanto, no permites que la página se muestre en caché.

- **Noodp / Odp:** se utiliza para indicarle al buscador que debe o no, mostrar el título y descripción de la página idénticos a los que se encuentra en el Open Directory Project. En algunas ocasiones determinados buscadores muestran como título de una web los que se han publicado en el ODP. El

Open Directory Project (también conocido como DMoz) es un proyecto colaborativo multilingüe, en el que editores voluntarios listan y categorizan enlaces a páginas web.

- **Noydir / Ydir:** esencialmente es lo mismo que el operador anterior, salvo que es para que no se pueda, o si, mostrar la descripción y título que aparece en el directorio de Yahoo.

También se puede utilizar el término **all** para referirse a todas las acciones permitidas por defecto (index, follow) y el término **none** para indicar las contrarias (noindex, nofollow).

En el caso de que la etiqueta robots presente instrucciones contradictorias (index, noindex), el crawler será el encargado de elegir la acción que desee tomar.

Además existen otra serie de operadores que no son muy utilizados ya que no son validados por todos los buscadores. Algunos sólo funcionan para los principales o incluso sólo para Google, como el operador **Unavailable_After,** el cual indica a Google que a partir de una fecha determinada la página debe dejar de ser indexada.

4.4.- Analizando robots.txt de algunas páginas

En este apartado realizaremos un análisis de los ficheros robots.txt de algunas páginas determinadas. Estos ficheros pasarán por el validador http://tool.motoricerca.info/robots-checker.phtml para determinar si están o no bien construidos.

Analyzing file http://www.google.com/robots.txt

Line 4	Disallow: /groups
Line 5	Disallow: /images
Line 6	Disallow: /catalogs
Line 7	Allow: /catalogs/about Unknown command. Acceptable commands are "User-agent" and "Disallow". A robots.txt file doesn't say what files/directories you can **allow** but just what you can **disallow**. Please refer to Robots Exclusion Standard page for more informations.
Line 8	Allow: /catalogs/p? Unknown command. Acceptable commands are "User-agent" and "Disallow". A robots.txt file doesn't say what files/directories you can **allow** but just what you can **disallow**. Please refer to Robots Exclusion Standard page for more informations.
Line	----------
Line 97	Disallow: /books/
Line 98	Disallow: /bkshp?*q=* The "*" wildchar in file names is not supported by (all) the user-agents addressed by this block of code. You should use the wildchar "*" in a block of code exclusively addressed to spiders that support the wildchar (Eg. Googlebot).
Line 99	Disallow: /books?*q=* The "*" wildchar in file names is not supported by (all) the user-agents addressed by this block of code. You should use the wildchar "*" in a block of code exclusively addressed to spiders that support the wildchar (Eg. Googlebot).

Ilustración 8. Validación de robots.txt de Google.com

Como se puede observar en la imagen, el robots.txt de Google presenta algunos errores (marcados en rojo). Al encontrar la directiva **Allow** en el fichero el validador nos muestra el siguiente error:

“Unknow command. Acceptable commands are User-agent and Disallow.
A robots.txt file doesn’t say what files/directories you can allow but just what you can disallow. Please refer to Robots Exclusion Standard page for more information.”

Esto nos indica que el comando **Allow** no es un comando conocido. Además nos indica que revisemos el Estándar de Exclusión de Robots para obtener más información. Si miramos el estándar de 1994 podemos observar que éste indica que el comando **Allow** no está permitido.

Parece ser que algunos robots como Google y Bing evalúan el comando **Allow** de una forma diferente:

- En el caso de Google, se evalúan primero todos los **Allow** y, una vez indexados, se encarga de todo excepto lo marcado por **Disallow**.
- Sin embargo Bing aplica la directiva más específica:
  - **Disallow: /articulo** & **Allow: /articulo/enero ->** No se indexará el contenido de articulo pero sí el de enero.
  - **Allow: /articulo** & **Disallow: /articulo/enero ->** El contenido de articulo se indexa, pero no el de enero.

No obstante, como el estándar indica que se aplica el primer match (la primera regla de coincidencia), puede darse el caso de que algunos robots se encuentren **Allow: /** & **Disallow: /articulo** y como la primera coincidencia coincide será la válida, por lo tanto, se indexará todo.

Por otra parte, si añadimos a esto que el comando **Allow** no está permitido en el estándar de 1994, puede darse la situación de encontrar un comportamiento distinto para cada tipo de robot, lo que nos lleva a preguntarnos: ¿Realmente conseguimos que no todos los buscadores indexen nuestros datos?

Otro de los errores que aparece en el robots.txt de Google es el que indica:

“The “*” wildchar in file names is not supported by (all) the user-agents addressed by this block of code. You should use the wildchar “*” in a block of code exclusively addressed to spiders that support the wildchar (Eg. Googlebot).”

Este mensaje nos indica que el asterisco * no es reconocido por todos los robots por lo que sólo debería usarse con robots que lo soporten. Al igual que en el caso del comando **Allow** existen distintos comportamientos sobre el uso del asterisco dependiendo del tipo de robot que lo indexe.

Analyzing file http://www.uc3m.es/robots.txt

The following block of code contains some errors. Please, remove all the reported errors and check again this robots.txt file.	
Line 1	`User-Agent: *` Although commands are not case sensitive, we advise you to write exactly "User-agent", that is all lowercase except for the capitalized "U".
Line 2	`Disallow: /portal/page/portal/seccion_dept_organizacion_empresas/catedrabancaja/publicaciones /Exploring%20corporate%20entrepreneurship%20in%20privatized%20firms.pdf`
Line 3	`Allow: /` Unknown command. Acceptable commands are "User-agent" and "Disallow". A robots.txt file doesn't say what files/directories you can **allow** but just what you can **disallow**. Please refer to Robots Exclusion Standard page for more informations.
Line 4	

Ilustración 9. Analizando el fichero robots.txt de uc3m.es

Al igual que en el caso del fichero robots.txt de la página anterior, el correspondiente a la Universidad Carlos III de Madrid también presenta el error de la directiva Allow. Sin embargo, aparte de este error encontramos uno en la primera línea del fichero, en la que se determina a qué robots afectan las directivas. El fallo que nos indica el evaluador es:

“Although commands are not case sensitive, we advise you to write exactly “User-agent”, that is all lowercase except for the capitalized “U”.”

Este mensaje nos indica que hay que escribir el comando **User-agent** correctamente, es decir, poniendo en mayúscula la primera letra y todo lo demás en minúscula.

4.4.1.- Precaución con los ficheros robots.txt

A pesar de que los ficheros robots.txt pueden ayudarnos a que los robots no indexen el contenido que nosotros no queramos que quede almacenado en los motores de búsqueda, un mal uso de ellos puede otorgar demasiada información valiosa para los hackers y otros curiosos.

Esto sucede porque muchas veces dichos ficheros contienen una gran cantidad de detalles en los comandos **Disallow** llegando a poner incluso el nombre del fichero completo como ocurre en la página oficial de los Grammy:

```
User-agent: *
Crawl-delay: 10
# Directories
Disallow: /includes/
Disallow: /misc/
Disallow: /modules/
Disallow: /profiles/
Disallow: /scripts/
Disallow: /themes/
# Files
Disallow: /CHANGELOG.txt
Disallow: /cron.php
Disallow: /INSTALL.mysql.txt
Disallow: /INSTALL.pgsql.txt
Disallow: /install.php
Disallow: /INSTALL.txt
Disallow: /LICENSE.txt
Disallow: /MAINTAINERS.txt
Disallow: /update.php
Disallow: /UPGRADE.txt
```

Ilustración 10. Fichero robots.txt de Grammy.com

Otros ficheros por el contrario, incluyen directorios que pueden ayudar a encontrar zonas internas de la página web dedicada a la administración de la misma, lo que facilita el acceso a estas zonas y anuncia su presencia a posibles hackers. Así ellos también pueden establecer algunos software empleados en el sitio web buscando “huellas típicas” en el robots.txt. Esto sucede por ejemplo en la página de la Real Federación Española de Karate:

```
User-agent: *
Disallow: /administrator/
Disallow: /cache/
Disallow: /components/
Disallow: /editor/
Disallow: /help/
Disallow: /images/
Disallow: /includes/
Disallow: /language/
Disallow: /mambots/
Disallow: /media/
Disallow: /modules/
Disallow: /templates/
Disallow: /installation/
```

Ilustración 11. Fichero robots.txt de Rfek.es

Como en el último ejemplo, una mala configuración del fichero robots.txt en el que se facilita ruta de la zona de administración, supuso que en 2007 un hacker accediera a la página del ministerio de vivienda de España y colgara en la página

principal una foto en la que se quejaba de la situación del país con respecto al tema de las viviendas [13].

En resumen, una mala configuración del fichero robots.txt puede suponer:

- Que el tiempo de indexación por parte del robot sea más lento y, por lo tanto, se penalice el rendimiento del servidor.
- Que no queden publicados documentos que sí queremos que lo sean.
- Que, por el contrario, queden publicados documentos a los que no deseamos permitir el acceso.
- Dar a conocer rutas y directorios a zonas de administración o versiones de software usadas en la página.

Puede ser que la mejor opción para configurar estos ficheros sea hacerlos más o menos público, sin dar a conocer las zonas más comprometidas de nuestra web. No es muy buena idea optar simplemente por la directiva **Disallow:** *, con la idea de no dejar en los buscadores ninguna información, ya que en el caso de encontrar esto, se pasa a buscar en los buscadores internos.

Los administradores deben comprobar el correcto funcionamiento de los robots.txt, ya que de no hacerlo, pueden estar definiendo directivas que no cumplan con su misión. Vamos a ver el ejemplo del fichero robots.txt de RTVE:

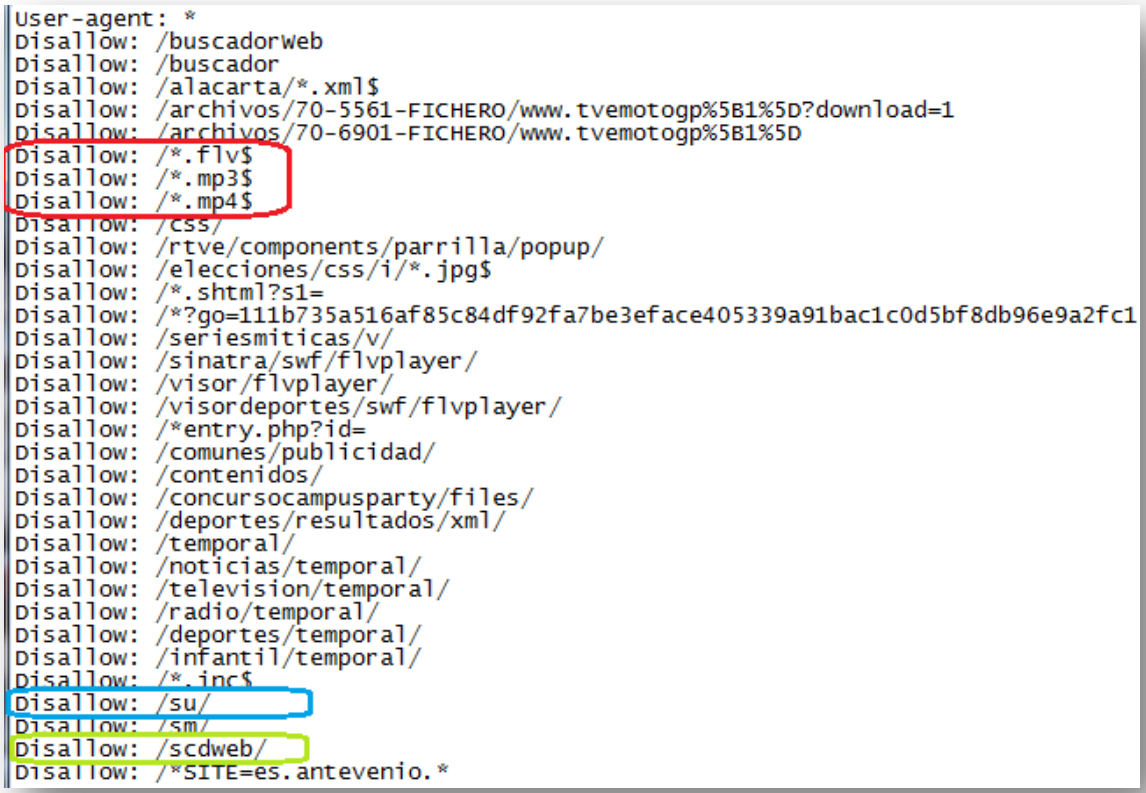

Ilustración 12. Fichero robots.txt de Rtve.es

Como podemos ver en la imagen, rodeado por un círculo rojo vemos directivas que el administrador ha establecido para prohibir el acceso a los robots

a los ficheros con extensión .flv, .mp3 y .mp4 entre otros. Sin embargo, si buscamos en Google alguno de estos ficheros mediante el comando **site:** rtve.es **ext:** flv.

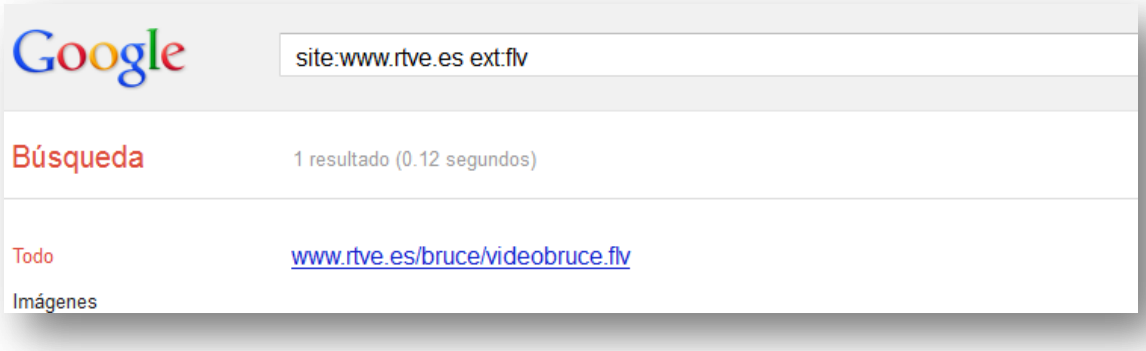

Ilustración 13. Archivo con extensión "prohibida"

También realizaremos una comprobación sobre uno de los directorios que supuestamente también están prohibidos para los robots (el marcado en verde en la imagen). Como podemos observar en la siguiente imagen, al realizar esta búsqueda también obtenemos un resultado:

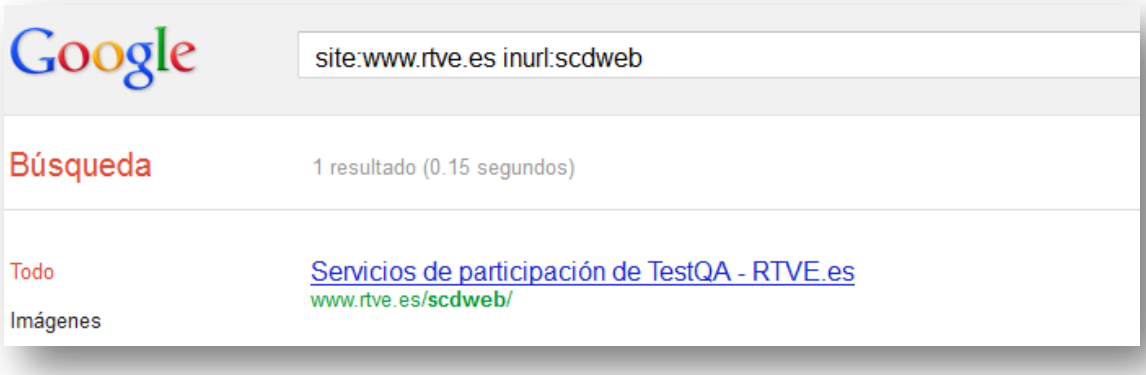

Ilustración 14. Ruta "prohibida" en RTVE

4.4.2.- Ficheros robots.txt hackeados

Puesto que, como hemos dicho en el apartado anterior, una mala configuración de un fichero robots.txt puede dar bastantes facilidades de acceso a zonas restringidas de un sitio web a hackers o curiosos, en este apartado pasaremos a buscar información en la red sobre ficheros robots.txt que hayan sido vulnerados. Esta búsqueda podemos realizarla porque muchas de estas personas dejan su "huella" en los robots.txt para que quede constancia de su intromisión.

Para realizar esta búsqueda utilizaremos los comandos **inurl:robots filetype:txt**, con los que indicaremos tanto que el término robots aparezca en la URL del sitio como que el fichero sea del tipo determinado, en este caso un documento de texto.

Si a los comandos anteriores le añadimos "defaced", el motor de búsqueda nos dará los resultados en los que el fichero robots.txt contenga esta palabra (defaced by = desconfigurado por). Realizando esta búsqueda en Google obtenemos lo siguiente:

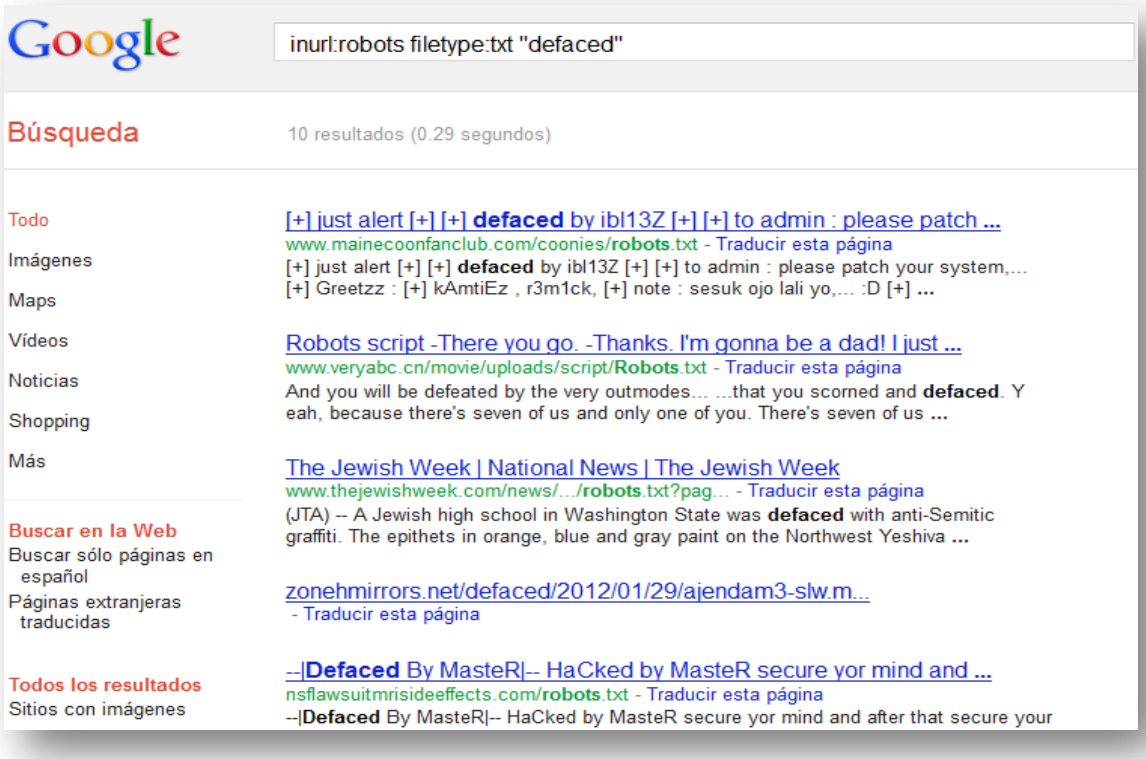

Ilustración 15. Búsqueda de robots.txt hackeados (I)

Si accedemos a algunos de los ficheros mostrados en la lista anterior, podemos observar robots.txt cuyo contenido ha sido modificado por los hackers. Hay casos, como el siguiente, en el que vemos que los hackers han hecho una advertencia al administrador del sitio web para que asegure su sistema.

```
[+] just alert [+]
[+] defaced by ibl13Z [+]
[+] to admin : please patch your system,...



[+] Greetzz :
[+] kAmtiEz , r3m1ck,
[+] note : sesuk ojo lali yo,... :D



[+] indonesianhacker team , indonesiancoder team [+]
```

Ilustración 16. Ejemplo de robots.txt hackeado (I)

En otros casos, podemos ver que a los hackers les gusta dejar su huella de una forma más “bonita”:

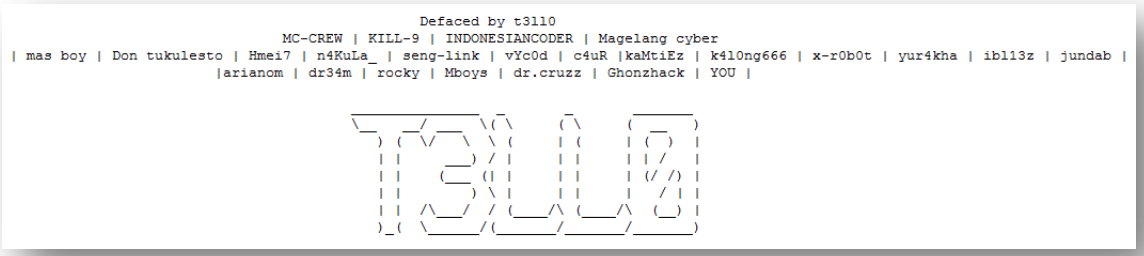

Ilustración 17. Ejemplo de robots.txt hackeado (II)

Vemos que mediante la búsqueda que hemos realizado anteriormente, no se obtienen demasiados resultados lo que nos parece un poco extraño, por lo que probaremos con otro término para ver si nos da nuevos enlaces. Utilizaremos el término “hacked”:

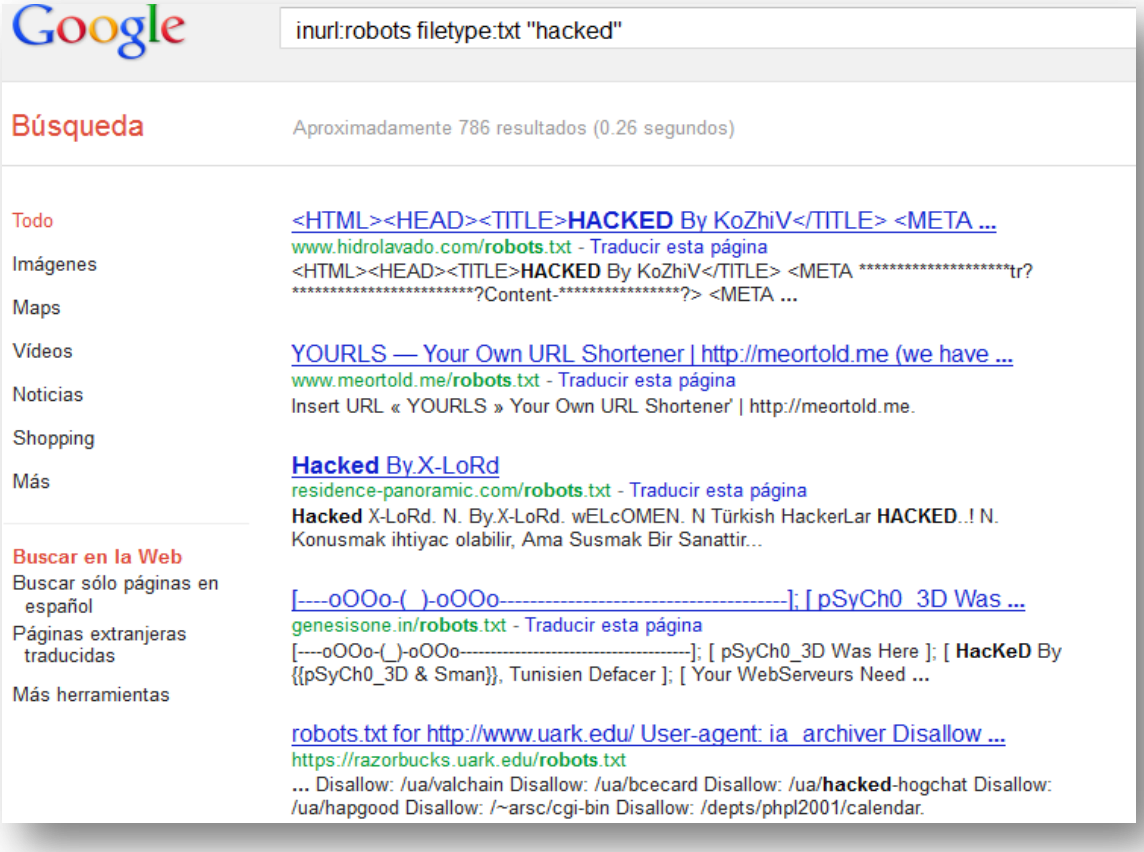

Ilustración 18. Búsqueda de robots.txt hackeados (II)

En esta ocasión podemos ver que ha aumentado considerablemente el número de resultados obtenidos, por lo que, como habíamos pensado en un principio, combinando determinados términos, podemos encontrar un gran número de ficheros robots.txt hackeados en la red.

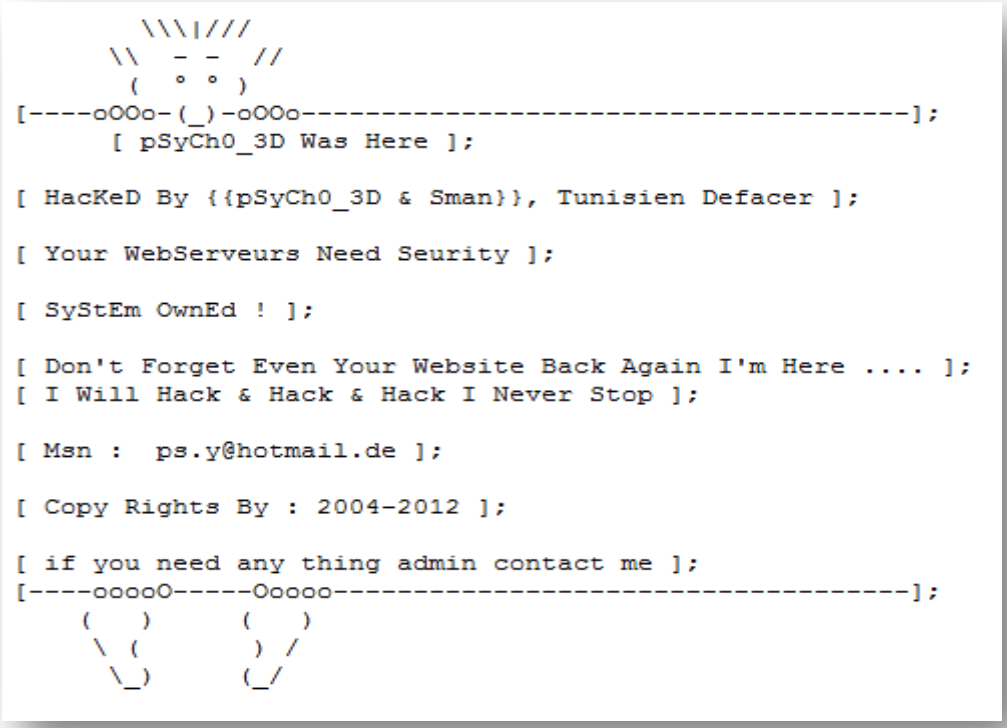

Ilustración 19. Ejemplo de robots.txt hackeado (III)

En la red podemos encontrar hackers tan “amables” que hasta dejan sus datos de contacto al administrador del sitio web por si éste necesita cualquier cosa.

5.- Proyecto Tor

5.1.- Definición

Tor (The Onion Router) es un software libre de encaminamiento de tráfico llamado *onionrouting* que tiene como fin hacer anónimas las comunicaciones en Internet contra la censura y el control. Fue diseñado originalmente como un proyecto del Laboratorio de Investigación Naval de los Estados Unidos. Estaba destinado a la Marina de los EE.UU. con la finalidad de proteger las comunicaciones gubernamentales. Actualmente se utiliza con una amplia variedad de propósitos tanto por militares como por periodistas, policías, usuarios normales y muchos otros.

Tor proporciona un túnel de comunicación anónimo diseñado para resistir a ataques de análisis de tráfico (*traffic analysis*). Por esta razón, Tor hace que sea posible que podamos realizar una conexión a un equipo sin que éste o ningún otro equipo sea capaz de determinar el número de IP de origen de la conexión.

En ocasiones se utiliza la combinación de Tor con Privoxy (programa que funciona como proxy HTTP diseñado para proteger la privacidad en la navegación en Internet) para acceder de forma segura y anónima a páginas web.

Tor permite a los desarrolladores de software generar nuevas herramientas de comunicación que incorporen características de privacidad. Suministra la base para una amplia gama de aplicaciones que permiten a las organizaciones y a los individuos compartir información sobre redes públicas sin comprometer su privacidad.

Los usuarios utilizan Tor para evitar ser rastreados por los sitios web, o para conectarse a sitios de noticias, servicios de mensajería instantánea o parecidos cuando han sido bloqueados por sus proveedores locales de Internet. Los usuarios pueden publicar sitios web y otros servicios sin necesidad de revelar la ubicación de los mismos por medio de los servicios ocultos de Tor. También es usado para la comunicación que se considera sensible socialmente, como pueden ser foros y salas de chat para supervivientes de violación y abuso, o personas con determinadas enfermedades.

Por otra parte, los periodistas utilizan Tor para conseguir una forma más segura de comunicarse con sus confidentes y disidentes, así como las organizaciones no gubernamentales (ONGs) lo usan para permitir que sus trabajadores puedan conectarse a sus sitios web cuando están en países extranjeros sin la necesidad de notificar a todo el mundo que está trabajando con esa organización.

Una de las cosas que hacen tan seguro Tor es la variedad de gente que lo utiliza. Te esconde entre los distintos usuarios de la red, por lo que, cuanto más diversa y mayor sea la red de Tor, mayor será el anonimato y, por lo tanto, estarás más protegido.

Aun teniendo en cuenta todo lo mencionado anteriormente, es importante aclarar que Tor no es 100% fiable en lo referente al cifrado de información. Está diseñado para asegurar el anonimato del usuario de manera que no se pueda seguir la información que envía para llegar hasta él. A su entrada, la red Tor cifra la información y la descifra a la salida de la misma, por lo que no es posible saber quién envió dicha información. A pesar de esto, el propietario de un servidor de salida puede ver toda la información cuando es descifrada antes de llegar a Internet, por lo que puede acceder a esa información aunque no le sea posible conocer el emisor de la misma.

Se recomienda utilizar algún sistema de cifrado como SSL para conseguir el anonimato en Internet y, además, la seguridad de que nadie consiga acceder a la información que está enviando. Asimismo, los desarrolladores de Tor recomiendan a los usuarios bloquear las cookies y los plugins Java, ya que éstos pueden averiguar la dirección IP del emisor. También se recomienda deshabilitar el historial de páginas web visitadas para tener mayor seguridad en el ámbito de atacantes con acceso físico a la máquina, como los que se encuentran en el mismo edificio que el usuario.

5.2.- Funcionamiento

A continuación, podemos ver una imagen con el aspecto básico que presenta la red Tor:

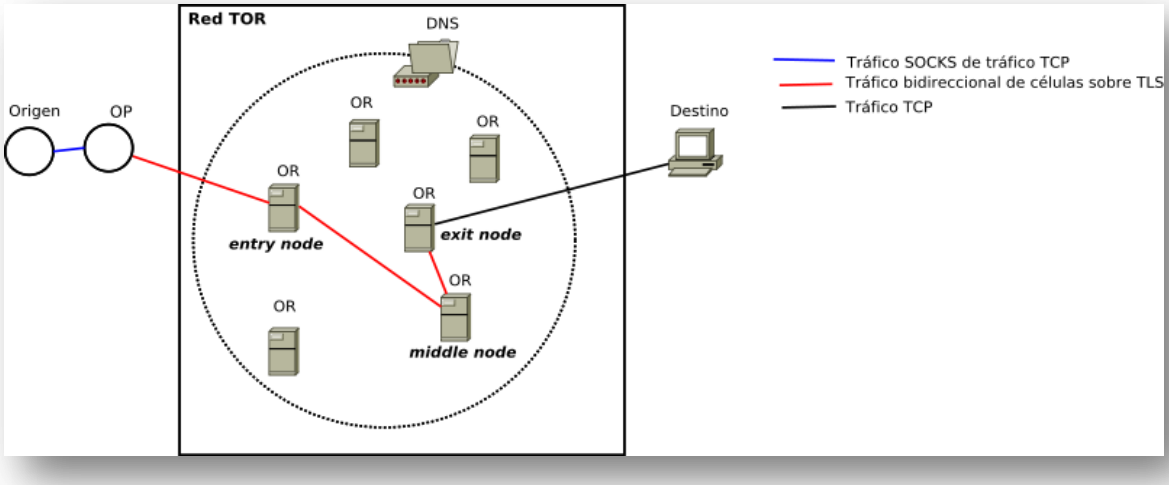

Ilustración 20. Componentes de la red TOR

La red se compone de una serie de nodos que se comunican entre sí mediante el protocolo TLS sobre TCP/IP lo que le hace mantener la información de una forma secreta e íntegra desde un nodo a otro. Se pueden diferenciar dos tipos de entidades:

- **Nodos TOR** (representados en la imagen como OR de Onion Router) cuya función es actuar como encaminadores y en alguna ocasión incluso como

servidores de directorio DNS de un servicio de mantenimiento. Se mantiene una conexión TLS entre cada uno de los nodos OR.
El servicio de directorio establece una base de datos que asocia a cada OR una determinada información (router decriptor). Todos los OR y usuarios finales pueden acceder a dicha información, usándola con el fin de tener un conocimiento sobre la red. En el caso de tener pocos servidores de directorio, se correrá el riesgo de tener un punto cuyo fallo podría ocasionar el fallo de todo el sistema. Los OR que dan el servicio de directorio mantienen duplicada la información enviándosela de unos a otros por motivos de backup y latencia.
Podemos establecer una diferencia entre una serie de OR principales (autoridades de directorio) y otros OR secundarios cuya misión es hacer de cachés y backup (directory caches).
TOR distribuye una lista de algunos de los servidores de directorio con el propósito de facilitar la suscripción a la red (bootstrapping). En el servicio de directorio se protegen las entradas criptográficamente con firmas, y únicamente la información que procede de ORs aprobados será publicada en la base de datos. De esta forma se previenen ataques en los que alguien quiera añadir muchos nodos no confiables. No existe un sistema automático que se encargue de aprobar las ORs, sino que los administradores del servidor de directorio lo hacen de forma manual.
Cuando un OR se inicia, recoge un conjunto de datos que lo describen tanto a él como a su modo de funcionamiento y capacidades. Algunos ejemplos de este conjunto de datos son: la dirección IP, versión del software TOR, nombre para el usuario, sistema operativo, clave pública y políticas de salida (cómo puede funcionar el nodo si es el último nodo del circuito de datos). Toda esta información es publicada a través del servicio de directorio.

- Los usuarios finales ejecutan un software local denominado onion proxy (OP). Este software obtiene la información del directorio, construye una serie de circuitos aleatorios a través de la red y trata conexiones de aplicaciones del usuario. Los OP aceptan flujos TCP de aplicaciones de usuarios y las multiplexa a través de la red ORs.

5.2.1.- Necesidad del uso de TOR

El uso de TOR supone una protección contra una manera habitual de vigilancia en Internet conocida como “análisis de tráfico”. El uso del análisis de tráfico supone poder deducir quién está hablando a quién sobre una red pública, por lo que los usuarios pierden su privacidad. Muchas entidades analizan el tráfico de Internet de los usuarios para conocer el origen y el destino del mismo, lo que les permite estar al tanto de su comportamiento e intereses. Los motores de búsqueda, las redes sociales, los medios de prensa en línea y las empresas analizan el tráfico de los usuarios para su propio beneficio económico. A pesar de ello, no se ha hecho ninguna declaración política sobre estos análisis que despojan a los usuarios de su propiedad (así como de su privacidad) en beneficio de la

democracia comercial del consumo. Este conocimiento del tráfico puede incluso amenazar al trabajo de los usuarios y su integridad física revelando quién es y dónde se encuentra. Por ejemplo, en el caso de conectarte a un ordenador de tu empresa desde un país extranjero, estás revelando tu nacionalidad y afiliación profesional a cualquier persona que se encuentre vigilando la red, aun cuando la conexión esté encriptada.

5.2.2.- ¿Cómo funciona el análisis de tráfico?

Los paquetes de datos que se envían a través de Internet están formados por dos partes: una carga útil de datos y una cabecera que se usa para enrutar el paquete. Los datos están compuestos de cualquier cosa que se esté enviando, ya sea un correo electrónico, un archivo de video o una página web. Aunque el usuario encripte los datos que envía en sus comunicaciones, el análisis de tráfico aún continúa revelando mucha información sobre él, tanto de lo que está haciendo como, posiblemente, de lo que está diciendo. Esto se debe a que el análisis se centra en la cabecera, la cual revela el origen, destino, tamaño, tiempo y demás información.

Uno de los problemas básicos en cuestión de privacidad es que el receptor de las comunicaciones de un usuario, puede ver lo que éste envía mirando las cabeceras. Por esta razón, el receptor puede autorizar a los intermediarios, como los proveedores de servicios de Internet, a que conozcan esta información, e incluso, a veces, también pueden ofrecer dicha información a intermediarios no autorizados. Una manera fácil de analizar el tráfico de una red consiste en situarse en algún lugar entre el emisor y el receptor de la información, y mirar las cabeceras de los paquetes de información que se envían.

También existen unas formas más potentes de análisis de tráfico. Algunos atacantes espían múltiples zonas de Internet y utilizan técnicas sofisticadas de estadísticas para rastrear patrones de comunicación de diversas organizaciones e individuos. Contra estos atacantes no sirve de ninguna ayuda la encriptación de la información, ya que simplemente esconde el contenido del tráfico entre los internautas, no el de las cabeceras.

Para solucionar este problema de privacidad, se diseñó una red anónima distribuida:

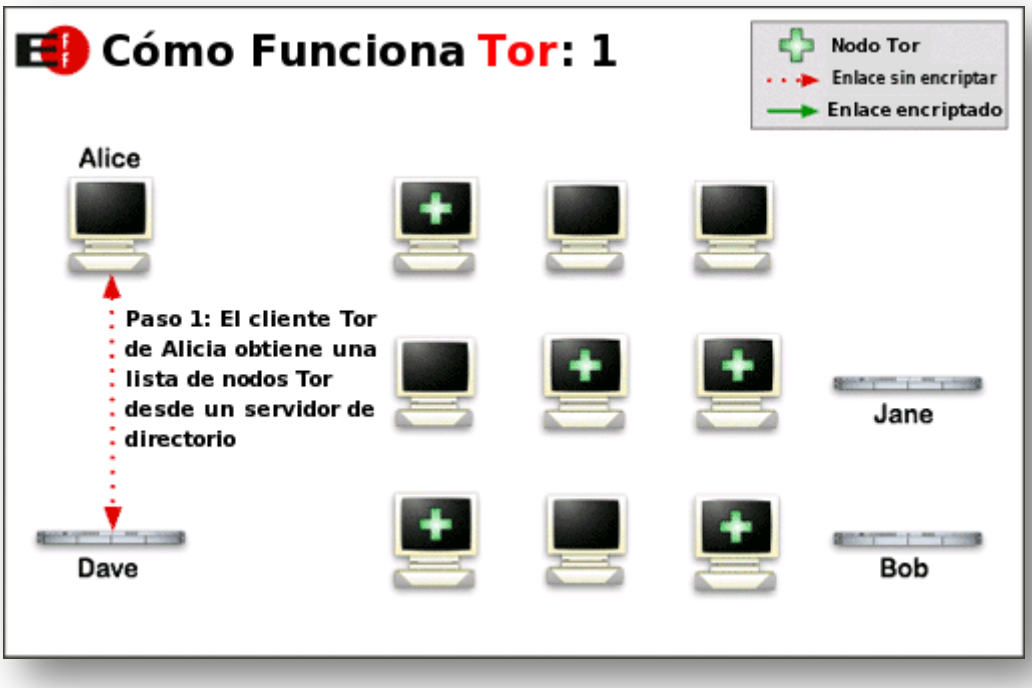

Ilustración 21. Funcionamiento de Tor (I)

Los riesgos de análisis de tráfico (tanto los sencillos como los sofisticados) son reducidos gracias a la ayuda de Tor, que distribuye las transacciones entre distintos sitios de Internet, razón por la cual ni un solo punto puede vincular al usuario con su destino. La idea es parecida a utilizar una curva, se establece una ruta difícil de seguir con la finalidad de despistar a alguien que esté siguiendo la actividad del usuario, además luego borra las huellas que haya podido dejar éste de forma periódica. Resumiendo, en vez de coger una ruta directa para la transmisión de paquetes de datos desde el origen al destino, éstos en la red Tor siguen un camino aleatorio a través de los routers que se encargan de camuflar las huellas de paso, para que ningún observador desde ningún punto de la red pueda saber de dónde vienen los datos ni a donde van dirigidos.

Para crear una ruta de red privada con Tor, el software del usuario o cliente construye de manera incremental un circuito de conexiones encriptadas a través de los routers de la red. Este circuito se extiende un tramo cada vez, y cada router que se encuentra a lo largo de ese camino, únicamente conoce el router desde el cual le están llegando los datos, y el router al que tiene que enviárselos. Cualquiera de los routers que se encuentran en la red no conocen el camino completo que un paquete de datos ha seguido. El cliente se encarga de establecer una serie de claves de cifrado para cada uno de los tramos que se encuentran en el circuito con el fin de asegurar que estas conexiones no puedan ser rastreadas por los routers a medida que los datos pasan por ellos.

El camino aleatorio que siguen los datos a través de la red Tor se puede ver en la siguiente imagen:

Ilustración 22. Funcionamiento de Tor (II)

Para finalizar, una vez que se ha establecido el circuito a seguir por los datos, muchos de ellos pueden ser intercambiados, y diferentes tipos de aplicaciones de software pueden ser implementados sobre la red Tor. Ya que cada router sólo puede ver un tramo del circuito, ningún espía ni ningún router comprometido puede usar el análisis de tráfico para asociar el origen de la conexión con el destino. Tor sólo funciona con flujos de información TCP y puede ser utilizado por cualquier aplicación que soporte SOCKS.

Por cuestiones de eficiencia, el software de Tor utiliza el mismo circuito generado para las conexiones que se originan en un intervalo de unos diez minutos más o menos. Después, a las nuevas solicitudes se les calcula un circuito nuevo con el fin de evitar que un individuo pueda asociar las primeras acciones de un usuario con las nuevas.

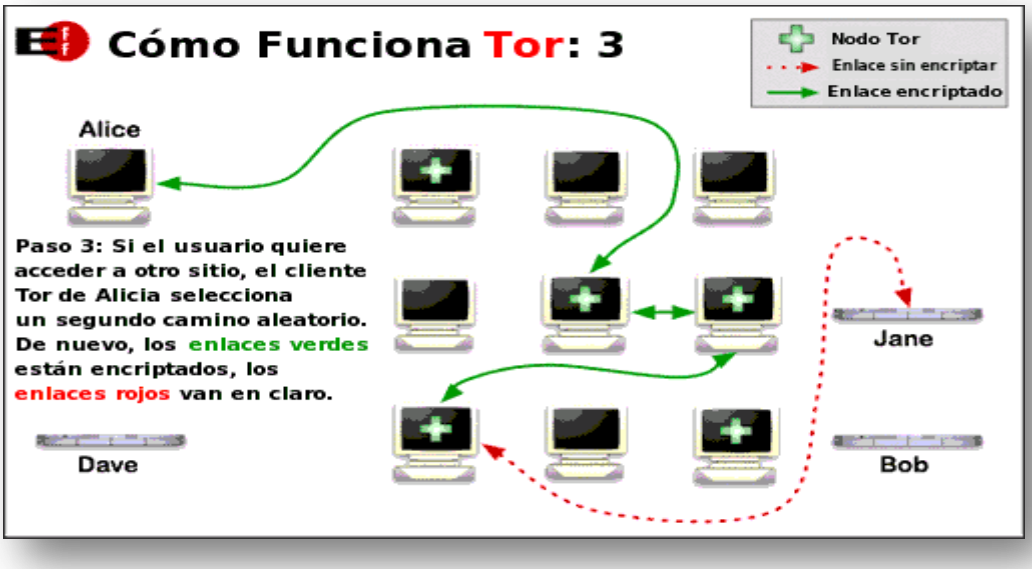

Ilustración 23. Funcionamiento de Tor (III)

5.2.3.- Protocolo de Tor

En este apartado definiremos a grandes rasgos el funcionamiento de Tor, dejando más claros los conceptos explicados anteriormente.

1. Partiendo de la información obtenida de su configuración y del servicio de directorio, el OP (onion proxy) elige un circuito a través del cual circularán los paquetes de datos. Por defecto, este circuito está compuesto por 3 nodos OR (onion router).

2. El OP establece unas claves de cifrado con cada OR perteneciente al circuito con el fin de proteger los datos durante todo el camino antes de realizar alguna transmisión. La obtención de las claves simétricas (AES-128), una para cada sentido de comunicación (Kf -> forward key, Kb -> backward key) parte del protocolo de establecimiento de claves Diffie-Hellman cuyo propósito es obtener una clave compartida desde la cual derivar dos claves simétricas. El circuito se construye desde el punto de entrada de la siguiente forma:
    1) Los mensajes para establecer las claves de la comunicación entre ORn y ORn+1 se realizan por petición del OP y enviando paquetes a través de los nodos OR1,…, ORn.
    2) En cada paso los mensajes se cifran con las claves de sesión negociadas o, en su defecto, con la clave ‘cebolla’ del host que recibe el dato.

3. A continuación se cifra el paquete que contiene la clave para el último OR del circuito.

4. Después, se hace lo mismo con el penúltimo y con los nodos restantes hasta llegar al primero.

5. Se envía el paquete obtenido al primer nodo del circuito. Este paquete está envuelto en varias capas de cifrado (razón por la que se usa la metáfora de cebolla).

6. El primer OR elimina su capa cifrada y envía el paquete al siguiente nodo.

7. Según vamos atravesando los nodos OR, cada uno de ellos irá eliminando su capa de cifrado correspondiente (la externa), por lo que ningún nodo OR perteneciente al circuito puede identificar el circuito completo ya que únicamente conoce los OR/OP anterior y posterior.

El primer nodo OR del circuito se conoce como ‘entry node’ y es el único que se comunica con el origen de la comunicación; a los nodos que se encuentran en el medio se les denomina ‘middle-node’; y el último nodo del circuito es el conocido como ‘exit-node’ y es el único que se comunica con el destino.

5.2.4.- Servicios Ocultos

Si bien la característica más conocida de Tor es proporcionar el anonimato a máquinas cliente finales, también puede ser usado para otorgar anonimato a servidores. A través del uso de la red Tor, los servidores anfitriones pueden ocultarse para evitar que su localización y quiénes lo usan sean conocidos.

Usando los “puntos de encuentro” de Tor, otros usuario de la red pueden conectarse a estos servicios ocultos sin dar a conocer su identidad ni conocer la del otro de la red. Esta funcionalidad de servicio oculto facilita la configuración de sitios web en los que los usuarios puedan publicar contenido sin tener en cuenta la censura. De esta forma nadie puede conocer quién ofrece un sitio determinado, así como los proveedores del servicio no pueden conocer quiénes publican en él.

Estos servicios utilizan una dirección .onion en lugar de un TLD existente. Aunque no existe un seguimiento de estos sitios, existen servidores que proporcionan direcciones útiles a los usuarios.

5.2.5.- Dominios .onion

Al sector Tor de la Internet invisible se le conoce como “onion”., mediante el uso de esta red se puede acceder a la web profunda “onion”. La entrada a la parte más invisible de la Internet Invisible se realiza a través de Onionland, o también (mal) denominada Darknet. Estos son sitios regulados bajo dominios del tipo .onion.

Los dominios .onion no son direcciones reales de Internet, y sólo se puede acceder a ellos a través de la red Tor. Estos dominios son difíciles de obtener, ya que están compuestos de una serie de caracteres y dígitos sin sentido generados aleatoriamente. El protocolo Onion Router que Tor establece como base para su funcionamiento, presenta los sitios más interesantes en cuanto a anonimato en la web invisible. La forma de acceder a ellos es conociendo su dirección IP y meterla en la barra de direcciones de Tor.

Existen algunos lugares de la red que clasifican y facilitan enlaces .onion. El más popular es la *hidden wiki*, que es un sitio web similar a la wikipedia en el que se almacenan estos enlaces clasificados en muchas secciones.

6.- La Web Invisible

6.1.- Tipos de Internet

Según Ricardo Fornas Carrasco no es posible considerar a Internet un único medio global de comunicación, ya que ningún internauta tiene acceso a todo Internet. Existen situaciones de inaccesibilidad a determinados contenidos y áreas de la red y conviene saber distinguirlos [14].

Se pueden distinguir tres tipos distintos de Internet atendiendo a la facilidad de recuperación de la información, ya sean: Internet global, Internet invisible e Internet oscuro.

6.1.1.- Internet Global

Red de información libre y gratuita a la que se puede acceder a través de la interconexión de ordenadores. Se accede a ella mediante programas navegadores, chats, mensajería o intercambio de protocolos (FTP, P2P).

6.1.2.- Internet Invisible

La Web Invisible, también conocida como Internet Invisible, Deep Web, Web oculta y otra serie de términos, es el conjunto de información presente en la web que, siendo accesible a través de Internet, no es posible encontrarla mediante los buscadores tradicionales. Esto se debe tanto a las limitaciones que tienen los spiders para acceder a todos los sitios web como el formato en el que se encuentran dichos sitios (bases de datos, páginas dinámicas, etc).

Contiene la información disponible en Internet pero a la que únicamente se puede acceder a través de páginas creadas dinámicamente tras realizar una consulta en una base de datos.

6.1.3.- Internet Oscuro

Se trata de los servidores y host a los que es imposible acceder a través de nuestro ordenador. La principal causa de esta inaccesibilidad se debe a las zonas restringidas con fines de seguridad nacional y militar, otras causas se deben a la incorrecta configuración de routers, servicios de cortafuegos y protección, servidores inactivos y, por último, servidores que han sido secuestrados para un uso ilegal.

6.2.- Clasificación de Internet según niveles

El total de Internet puede agruparse en ocho niveles:

- **Nivel 1 - Surface:** La superficie, es el nivel que todos los usuarios conocen y visitan frecuentemente. En este nivel encontramos páginas típicas como Youtube, Twitter, Bing, etc.

- **Nivel 2 – Bergie:** Este nivel incluye, entre otras páginas, los foros chan. En este nivel nada es ilegal, pero ya se recomienda tomar precauciones.

- **Nivel 3 – Deep:** En este nivel ya conviene usar medidas de protección. Se encuentra mucho material ilegal aunque también páginas de gran calidad y alto contenido informativo de varios temas.

- **Nivel 4 – Charter:** En este nivel es necesario utilizar programas que permitan el anonimato (como Tor) ya que el material que incluye es en gran parte ilegal, como información robada a empresas, atrocidades cometidas por el ser humano, etc.

- **Nivel 5 – Zion:** A este nivel es muy difícil acceder debido a la necesidad de hardware especial. Se mezcla la realidad con las leyendas, ya que poca gente conoce su contenido. Se dice que en este nivel se encuentran los archivos robados a los gobiernos, conspiraciones, experimentos secretos, etc. Zion está, a su vez, dividido en otros 3 nuevos niveles (a los que denominaremos niveles 6, 7 y 8).

- **Nivel 6:** Acceder a este nivel supone estar expuesto a muchos más problemas de los que se estaba en los niveles anteriores, incluso para los usuarios más expertos. Para poder acceder a él, es necesaria la computación cuántica.

- **Nivel 7:** Este nivel ofrece una gran cantidad de poder, ya que en él se encuentran todo tipo de códigos para evitar que la gente avance. Por esta razón, se define como una zona de Guerra, donde se encuentran usuarios que intentan acceder a este nivel y usuarios que intentan expulsar a otros del mismo. Hay información tan ilegal que incluso muchos gobiernos no tienen acceso a ella.

- **Nivel 8 – Primario:** Es el que controla Internet, no está controlado por ningún gobierno u organización y es imposible de acceder de forma directa. Este sistema es una anomalía descubierta en el año 2000 y se diferencia del nivel 7 mediante el código: "level 17 qantum t.r.001 levelfunction lock", prácticamente irrompible por nuestras computadoras. Nadie conoce este nivel y se piensa que quien consiga acceder a él podrá controlar el Internet a su antojo.

6.3.- Tipos de Información

Los tipos de información que podemos encontrar dentro de la Internet Invisible son los siguientes:

- **Bases de datos:** los buscadores únicamente permiten el acceso a las páginas principales, ya que las demás se generan dinámicamente. Las principales bases de datos que se encuentran son:
    - **Bases de datos bibliográficas:** en las que se incluyen los catálogos de bibliotecas y librerías, referencias bibliográficas, etc.
    - **Bases de datos alfanuméricas o a texto completo:** dentro de esta categoría se incluyen las obras de referencias como enciclopedias o diccionarios.
    - **Bases de datos referenciales:** incluyen directorios de empresas, organizaciones, legislación, etc.
- **Documentos en formatos no indizables:** inicialmente esta categoría incluía formatos del tipo pdf, doc, xls, y otros tipos, ya que los buscadores únicamente indizaban formato html. Actualmente algunos buscadores ya permiten la indización de este tipo de formatos.
- **Revistas electrónicas y archivos de documentos:** son invisibles a los motores de búsqueda e incluyen tanto las de pago (a las que se suele acceder mediante IP o palabra clave) como las gratuitas (a las que se accede a través de registro).
- **Páginas web no indizadas:** que son excluidas expresamente por el propietario utilizando algún protocolo de exclusión.
- **Páginas web con contraseñas:** a las que los buscadores no pueden acceder.
- **Información generada dinámicamente:** como es el caso de tablas estadísticas, resultados deportivos, cambios de moneda, mapas y planos, premios de lotería, tesauros, etc.

6.4.- Origen del término

En 1994, la doctora especializada en Internet Jill Ellsworth, usó el término “Internet invisible” para hacer referencia a la información que no era posible encontrar en los buscadores más comunes [15]. Por otra parte, Michael Bergman realizó un estudio en 2000 [8] en el que ratificaba la presencia de una “red profunda” cuyo tamaño era de aproximadamente 7500 terabytes frente a los 19

terabytes de la web visible, aquella parte de la web a la que se puede acceder a través de los buscadores convencionales.

Por otra parte, Lluís Codina opina que Internet Invisible “es un nombre claramente inadecuado para referirse al sector de sitios y de páginas web que no pueden indizar los motores de búsqueda de uso público como Google o Altavista”. Propone el término de web “no indizable”, que en su opinión es “un término mucho más adecuado, pero claramente alejado de la capacidad de sugeridora del término invisible” [16].

6.5.- Causas de su existencia

La Web Invisible apareció ante los ojos de muchos individuos que imaginaban que las fronteras de la web estaban más lejos de lo que realmente parecía mediante el uso de los motores de búsqueda. En esta web profunda podemos encontrar tanto grandes fuentes de información muy valiosa como información “oscura” que puede llevarnos a la cárcel.

La principal razón de la existencia de la Web Invisible es la imposibilidad de los buscadores de localizar e indexar el 100% de la información presente en la red. Si estos buscadores fueran capaces de acceder a toda la información de la red, desaparecería la parte “invisible” de Internet, lo cual no es posible porque siempre habrá páginas privadas.

Los spider que emplean los buscadores para recorrer las páginas de la web siguen los enlaces que presentan o se dirigen hacia ellas y después las almacenan en una base de datos propia, por lo que a la hora de buscar una página web realmente no la estamos buscando en la red, sino en la base de datos del buscador en cuestión.

Otra de las causas de la existencia de esta parte de Internet se debe a que las páginas dinámicas no son indizadas por los buscadores, por lo que no pueden aparecer en ellos. Además se presenta otro problema con el formato en el que se guarda la información. En su origen, los buscadores fueron creados para leer, indizar y descargar páginas HTML, por lo que cualquier otro formato de información se volvía invisible para ellos (como por ejemplo videos, imágenes, pdfs, etc). Muchos de los buscadores actuales no son capaces de recuperar estos archivos, aunque éste no es el caso de Google y Altavista que ofrecen algunas posibilidades para la búsqueda de este tipo de formatos. Que no se indicen este tipo de formatos supone un gran problema ya que, muchos documentos que contienen información valiosa son prácticamente inaccesibles a la mayoría de los usuarios, a pesar de estar publicados y disponibles en la web de forma pública.

En el caso de las bases de datos, es posible acceder a sus páginas principales porque están definidas en formato HTML, pero no se puede acceder al resto del sitio a través de los buscadores, lo que puede ocasionarnos una gran pérdida de información. Una posible solución a este problema sería crear interfaces de consultas que envíen una determinada consulta a diferentes bases de datos desde

el mismo sitio web. Este es el modelo que siguen los multibuscadores o metabuscadores.

Por otro lado, también existen páginas web que evitan ser indexadas por los spiders a través del uso de protocolos de exclusión. Este protocolo consiste en añadir etiquetas meta dentro del código fuente de la página o mediante el empleo de ficheros robots.txt.

6.6.- Tamaño

Según un estudio realizado por Michael K. Bergman en 2000, el tamaño de la Internet Invisible se estimaba en unos 7.500 TeraBytes de datos en unos 550.000 millones de documentos. Como comparación, en aquella época se estimaba que la Internet Visible ocupaba 167 TeraBytes. Esto supone que más de 250.000 sitios web pertenecen a la web profunda. Además, el contenido de la Biblioteca del Congreso de Estados Unidos se basaba en unos 3.000 Terabytes imposibles de acceder por medio de los motores de búsqueda. Estos datos se refieren únicamente a los sitios web en inglés o con caracteres europeos.

La web oculta va aumentando a medida que pasa el tiempo debido a dos razones principales: en primer lugar, las nuevas fuentes de datos suelen ser del tipo búsqueda/petición dinámica, ya que éstas tienden a ser más útiles que las páginas estáticas. En segundo lugar, los gobiernos de todo el mundo, han decidido consentir el acceso a través de Internet a sus documentos oficiales y registros.

Parece ser que los sitios pertenecientes a la Internet Invisible, reciben un 50% más de tráfico que las páginas web pertenecientes a la visible y, aunque no sean conocidos de forma pública, tienen un mayor número de sitios enlazados. Normalmente están compuestos de ámbitos más concretos, pero con contenidos más profundos y detallados. La razón de este mayor número de tráfico puede deberse a que la calidad de estos contenidos es entre 1.000 y 2.000 veces mayor que los contenidos de la web visible, debido a la exhaustividad de los mismos. En la siguiente tabla podemos ver una comparación de los contenidos de los dos tipos de web (visible e invisible):

Consulta	Internet Visible			Internet Invisible		
	Total	"Calidad"	Porcentaje	Total	"Calidad"	Porcentaje
Agricultura	400	20	5%	300	42	14%
Medicina	500	23	4.6%	400	50	12.5%
Finanzas	350	18	5.1%	600	75	12.5%
Ciencias	700	30	4.3%	700	80	11.4%
Derecho	260	12	4.6%	320	38	11.9%
TOTAL	2.210	103	4.7%	2.320	285	12.3%

Tabla 2. Recuperación de documentos de calidad: Internet Visible Vs. Internet Invisible. En: Bergman, Michael K. The Deep Web: Surfacing Hidden Value

El mismo estudio realizado por Michael K. Bergman indica que el 95% de los contenidos de la Internet Invisible son totalmente públicos, sin ningún tipo de coste y almacenado en su mayor parte en bases de datos especializadas. Además, la parte profunda de la web crece a un ritmo más alto que la parte visible.

En la siguiente imagen se muestra la tipología y distribución de los contenidos pertenecientes a la web invisible:

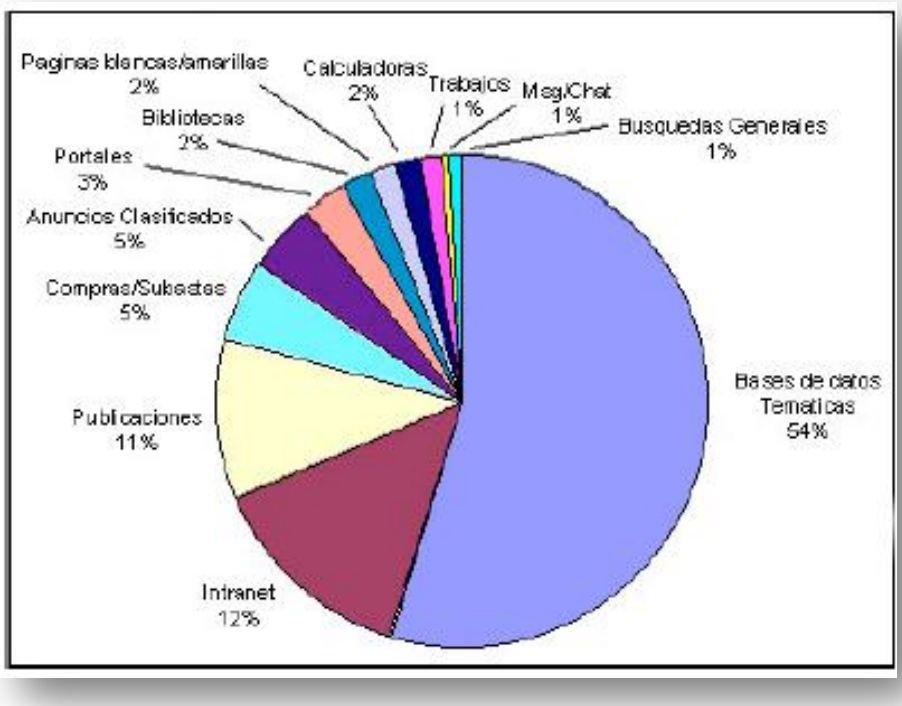

Figura 9. Distribución de los recursos en Internet Invisible.
En: Bergman, Michael K. The Deep Web: Surfacing Hidden Value.

Por otro lado, en la siguiente tabla se puede observar la cobertura de las distintas áreas temáticas contenidas en la web profunda. En ella se puede ver que la categoría de Humanidades es la que obtiene más cobertura que las demás, con su porcentaje del 13.5%. A continuación, se encuentra la relacionada con las noticias y medios de comunicación y en último lugar tenemos la relacionada con la agricultura.

Cobertura de Internet Profunda / Internet Invisible	
Humanidades	13.5%
Noticias, Media	12.2%
Ordenadores, Web	6.9%
Arte	6.6%
Negocios	5.9%
Salud	5.5%
Gente, Compañías	4.9%
Referencias	4.5%
Educación	4.3%
Empleo	4.1%
Ciencias, Matemáticas	4%
Estilo de vida	4%
Derecho, Política	3.9%
Gobierno	3.9%
Recreo, Deportes	3.5%
Viajar	3.4%
Compras	3.2%
Ingeniería	3.1%
Agricultura	2.7%

Tabla 3. Distribución de la cobertura por áreas temáticas.
En: Bergman, Michael K. The Deep Web: Surfacing Hidden Value.

6.7.- Instrumentos de búsqueda

Cada buscador, directorio o metabuscador está especializado en la búsqueda y selección de determinados contenidos de páginas web. Por esta razón, existe una variedad de mecanismos de búsqueda de información. En Internet Invisible, las herramientas existentes para la búsqueda de información son muy parecidas a las que los usuarios utilizan en la Internet Visible, diferenciándose en la capacidad de cada una para realizar búsquedas especializadas y selectivas en las distintas áreas que no pueden explorar los motores de rastreo.

A continuación se presenta una tabla comparativa en la que se pueden ver las diferencias entre las características de búsqueda para la web visible y la invisible de las principales herramientas de búsqueda: buscadores, directorios y metabuscadores.

	INTERNET VISIBLE	INTERNET INVISIBLE
Diseño	Diseño con lenguaje HTML y Flash.	Diseño con lenguajes SQL, XML.
Contenido	Macro RED con contenidos: públicos y privados.	Redes paralelas con contenidos: opacos, privados, propietarios y

		ocultos.
INSTRUMENTOS DE BÚSQUEDA		
Buscadores	Exploran, indizan y almacenan direcciones de sitios web en sus bases de datos. Buscan términos y palabras específicas. La exploración es periódica y automática. Calculan la relevancia de la información dependiendo de la frecuencia de las palabras y las páginas ligadas a otras. Consultas por palabras específicas. **Ejemplos:** Google, Bing, Tahoma.	Rastrean contenidos de bases de datos y contenidos como imágenes y documentos PDF. No indizan contenidos de páginas dinámicas. Consultas por palabras y frases. **Ejemplos:** InvisibleWeb.com, LexiBot, Lycos.
Directorios	Índices temáticos ordenados por categorías. Registro y clasificación manual. Contenidos selectivos. La heterogeneidad de las categorías se debe a razones: técnicas, servicios prestados, económica, cuota de inscripción. Consultas por: categoría. **Ejemplos:** Yahoo, LookSmart.	Índices temáticos especializados ordenados por categorías. Creadas y mantenidas por bibliotecarios o investigadores. Consultas por: categorías. **Ejemplos:** infomine, librarían Internet Index, buscopio.
Metabuscadores	No tienen mecanismos propios de búsqueda, sino que buscan en los buscadores existentes. La búsqueda puede ser: agrupada (en varios buscadores) o individual (en un buscador). Especifican las fuentes. Posibilidad de búsqueda avanzada. Consultas por: palabras. **Ejemplos:** Dogpile, InfoSpace, MetaCrawler.	Se conecta con buscadores especializados en la web profunda. Posibilidad de búsqueda avanzada. Consultas por: palabras y frases. **Ejemplos:** CompletePlanet, Topic Hunter, profusion.

Tabla 4. Análisis comparado de características de las herramientas de búsqueda de contenidos en Internet visible e Invisible.

6.8.- Clasificación de la Internet Invisible

En un estudio realizado en 2001, Sherman y Price clasificaron la Internet Invisible en cuatro categorías distintas: la web opaca (the opaque web), la web privada (the prívate web), la web propietaria (the proprietary web) y la web realmente invisible (the truly invisible web) [17].

6.8.1.- Web Opaca

Esta categoría está formada por los archivos que podrían aparecer en los motores de búsqueda pero que no lo hacen por alguna de las siguientes razones:

- **Extensión de la indización:** los buscadores no indizan todas las páginas de un sitio web por razones económicas.

- **Frecuencia de la indización:** aunque cada día se crean, modifican o eliminan páginas web, la indización realizada por los motores de búsqueda no sigue el mismo ritmo, razón por la cual hay páginas existentes en la web que aún no han sido indexadas y, por lo tanto, forman parte de la web invisible.

- **Número máximo de resultados visibles:** no todas las páginas existentes en la web, aun siendo indexadas por los buscadores, aparecen en la lista de resultados generada por los mismos, ya que éstos limitan el número de documentos mostrados (entre 200 y 1000 documentos).

- **URL’s desconectados:** los grandes buscadores actuales presentan la lista de resultados en orden de relevancia de los documentos según éstos hayan sido ligados en otros y basándose en el número de veces que aparecen referenciados. Si un documento no se encuentra enlazado por ningún otro, es posible que éste no sea descubierto, ya que no habrá sido indizado.

6.8.2.- Web Privada

Esta categoría está formada por las páginas web que podrían aparecer en los motores de búsqueda pero que éstos no las indizan debido a que son excluidas intencionadamente por algunas de las siguientes razones:

- El propietario del sitio web pretende mantener una página sin ser enlazada desde ningún otro sitio dentro de su propio dominio, de esta forma un usuario no puede encontrar dicha página navegando dentro de esa web. Esta técnica no es muy efectiva ya que aunque el propietario intente mantener esa página oculta, puede aparecer en algún lugar su enlace y, por lo tanto, será referenciada.

- La página está protegida mediante el uso de contraseñas (passwords). Se puede hacer mediante dos formas principales: la primera es utilizando el archivo .htaccess (esta forma se utiliza cuando no se tiene acceso al servidor), y la segunda es utilizando el panel de administración (cuando tienes acceso al servidor).

- La página contiene un archivo robots.txt que no permite la indización de la misma o de partes de la misma.

- En la página aparece una etiqueta “noindex” que le indica a los robots de los buscadores que no deben indizar esa página.

- El propietario ha bloqueado la URL de la página en Google Webmaster Tools. Esta herramienta permite eliminar la página de Google una vez que haya sido indexada, pero no impide su indexación.

6.8.3.- Web propietaria

En esta categoría se incluyen las páginas a las que los usuarios solo pueden acceder a su contenido mediante el registro en las mismas, ya sea de forma gratuita o pagada. Como se indicó en el punto anterior, un estudio realizado por Michael K. Bergman indica que el 95% de los contenidos de la web son de acceso público y gratuito.

6.8.4.- Web Realmente Invisible

Esta categoría está compuesta por aquellas páginas que no pueden ser indizadas por los buscadores debido a limitaciones técnicas de los mismos, como por ejemplo:

- Páginas web que contienen documentos en formatos pdf, PostScript, Flash, Shockwave, programas ejecutables y archivos comprimidos.

- Páginas dinámicas. Aquellas generadas partiendo de los datos que mete el usuario.

- Información almacenada en bases de datos relacionales. Esta información no puede ser extraída a menos que se haga una petición específica sobre ella. Además, se añaden otras dificultades como la estructura y diseño de las bases de datos y los procedimientos de búsqueda existentes.

Sin embargo, a lo largo de estos años los grandes buscadores como Google han ido desarrollando algoritmos nuevos que le permiten rastrear algunos formatos de archivos, documentos y bases de datos mencionados anteriormente que antes quedaban excluidos.

6.9.- La Web Invisible en la Actualidad

Atendiendo a la clasificación realizada en el apartado anterior, en este apartado vamos a analizar la situación actual de la información correspondiente a cada una de las categorías mencionadas anteriormente.

6.9.1.- Web Opaca

Con el paso de los años, los buscadores y directorios de la web van mejorando tanto su estrategia de búsqueda como los mecanismos para la misma. Se ha conseguido un aumento de la eficiencia de los buscadores bastante considerable, lo que nos permite obtener un mayor número de resultados en las búsquedas de los que se obtenían antiguamente. Mediante la evolución de estas herramientas de búsqueda será más necesario orientar a los usuarios en las estrategias de búsqueda y en el uso y aprovechamiento de los recursos localizados, y se dejará un poco de lado la elaboración de guías o concentradores de recursos.

En cuanto a las razones que se comentaron en el apartado anterior perteneciente también a la web opaca, podemos ver que actualmente se encuentra en la siguiente situación:

- Se mantiene la práctica de no indizar todas las páginas de un sitio web. Aún podemos encontrar por ejemplo la referencia a una base de datos publicada en un sitio web y no encontrar la referencia a la página de acceso directo a la base de datos de ese sitio.

- En algunos buscadores ha aumentado la frecuencia de indización ya que ahora el número de páginas creadas, modificadas o eliminadas al día, es bastante mayor del que era hace unos años. También algunos incluyen distintas formas de indización dependiendo de los recursos indizados. Los robots visitan con más frecuencia las páginas que, por su contenido, tienden a variar más que las páginas que mantienen un contenido más estable.

- El número máximo de resultados mostrados no es un problema a la hora de encontrar información ya que la mayoría de los buscadores actuales muestran los resultados ordenados por relevancia, por lo que al realizar una búsqueda, siempre aparecerán en las primeras posiciones los documentos más relevantes. Los usuarios deben ser conscientes de que normalmente son más efectivos los motores de búsqueda a la hora de realizar búsquedas específicas, mientras que a la hora de realizar búsquedas temáticas son más eficaces los directorios.

- Una forma de resolver el problema de las URL’s desconectadas podría ser obligando a registrar cualquier página que se colgara en la web. Sin embargo, debido a la gran descentralización actual de Internet no parece que vaya a solucionarse en un futuro inmediato.

6.9.2.- Web Privada

Esta categoría no ha tenido muchas modificaciones desde su inicio ya que, al tratarse de páginas excluidas conscientemente, no hay mecanismos que puedan aplicarse para que sean visibles. No presenta una gran pérdida de información valiosa ya que normalmente se trata de páginas que no tienen gran utilidad.

Por otra parte, los archivos robots.txt son bastante útiles para evitar que los robots caigan en “agujeros negros” (que entren en procesos circulares interminables), por lo que la eficiencia de los mismos no disminuye.

6.9.3.- Web Propietaria

El contenido perteneciente a esta categoría ha aumentado con el paso de los años debido, en gran medida, a la aparición de las redes sociales, que necesitan el registro de los usuarios para acceder a las mismas.

Por otra parte, el mayor contenido de la web está originado por Estados Unidos y con páginas en inglés, por lo que cabe esperar que se encuentren más páginas en la Web Invisible originadas en países diferentes a los Estados Unidos y que presenten un idioma que no sea el inglés.

6.9.4.- Web Realmente Invisible

Como se indicó en el apartado anterior perteneciente a la web realmente invisible, con el paso de los años los grandes buscadores han evolucionado y permiten realizar búsquedas por materiales o formatos especiales que antes no eran capaces. Por ejemplo, Altavista permite realizar búsquedas de imágenes, audio y video. Google da la posibilidad de utilizar mecanismos de búsqueda para encontrar imágenes, videos, noticias, etc. Por otro lado, HotBot permite realizar búsquedas por distintos formatos, añadiendo a los ya mencionados anteriormente (imágenes, audio, etc) otros nuevos formatos como pdf, script, shockwave y flash.

Para que sean posibles las búsquedas de estos formatos especiales, se realiza una catalogación textual de los mismos. Por ejemplo, los formatos pdf, flash, etc, se pueden encontrar ya que existen directorios que almacenan estos tipos de archivos. Esto indica que el principal medio por el que podemos realizar las búsquedas es por el texto. En el caso de que queramos localizar imágenes, éstas deberán estar clasificadas en la base de datos, lo que implica un proceso manual para elaborar la clasificación. Para devolver los resultados más satisfactorios posibles a la hora de realizar una búsqueda de imágenes, previamente una persona habrá tenido que clasificar los distintos tipos de imágenes, para que así se pueda devolver al usuario la imagen que más se asemeje a la consulta realizada.

Aunque las máquinas actuales son bastante eficaces a la hora de recuperar información textual, esto no impide que haya que catalogar y clasificar los distintos recursos. Por esta razón, el número de archivos con ese formato que puedan recuperar los buscadores es limitado. Si una página contiene una imagen que no tiene asignada ningún tipo de información textual, no podrá ser recuperada por los buscadores más que por su extensión (.jpg, .png, etc).

6.9.5.- Conclusiones

Lo que realmente aún en día continua siendo invisible en la web son:

- URL’s desconectadas.
- Páginas no clasificadas que contienen principalmente archivos pdf, PotScript, Flash, Shockwave, ejecutables y comprimidos.
- Páginas no clasificadas que contienen principalmente imágenes, audio y videos.
- Contenido de las bases de datos relacionales.
- Contenido generado en tiempo real.
- Contenido generado dinámicamente.

Aún así:

- Algunos buscadores actuales ya son capaces de recuperar archivos pdf e imágenes, aunque de forma limitada.
- Es relativamente sencillo llegar hasta la entrada de bases de datos con contenido importante.
- Existen buscadores avanzados que pueden lanzar búsquedas simultáneas en distintas bases de datos a la vez, aunque la mayoría son de pago.
- El contenido generado en tiempo real pierde validez muy rápidamente, salvo en el caso de los análisis históricos.
- Es relativamente sencillo llegar a los servicios que ofrecen información en tiempo real.
- El contenido generado dinámicamente sólo interesa a ciertos usuarios.
- Es relativamente sencillo llegar a los servicios que ofrecen contenido generado dinámicamente.

6.10.- Accediendo a la Internet Invisible

Existen diversas formas de acceder a la Internet Invisible, aunque las principales son la visión crawling y la visión metasearch. Cuando se accede mediante crawling, lo que se hace es realizar consultas preestablecidas sobre los distintos formularios de búsqueda para acceder a la información y así poder indexar las páginas resultantes.

Por ejemplo, un acceso de tipo crawling sería realizar una consulta preestablecida como puede ser el título de un libro “The Hypnotist”. El crawler contiene una base de datos de enlaces web, recorre cada uno de ellos y, cuando encuentra un formulario, tiene que identificar el campo que se corresponde con el título del libro en este caso. Lo que hace para identificar este campo es utilizar técnicas heurísticas en las que descubre el texto que designa a cada campo. Una vez ha descubierto el campo que tiene que rellenar, mete el valor de la consulta (el título del libro) y realiza el envío del formulario.

Por otra parte, para acceder a la Internet Invisible mediante metasearch, el proceso que se sigue es combinar todos los formularios de búsqueda pertenecientes a un mismo dominio temático en un mismo formulario que tiene que ser rellenado por el usuario. Estos formularios unificados pueden pertenecer a venta de libros, DVDs o música. Cuando el usuario ha rellenado el formulario unificado, se rellenan los formularios objetivo con los valores especificados previamente. Estos formularios son consultados en tiempo real y se ofrece las respuestas de cada uno al usuario.

6.11.- ¿Qué podemos encontrar en la Internet Invisible?

En la Internet Invisible se puede encontrar una gran cantidad de información, el problema es saber encontrarla. La mayor parte de los usuarios cree que la Internet Invisible sólo almacena contenido ilegal e inmoral, pero esto no es cierto, ya que se puede encontrar información bastante útil y de gran calidad. Podemos encontrar desde bibliotecas con mucho material, revistas, diccionarios, expedientes y archivos clasificados, hasta un gran número de actividades ilegales, como páginas de pedofilia, venta de drogas, construcción de bombas, etc.

Se recomienda mantenerse alejado de cualquier cosa que aparezca etiquetada como “chan”, “CP” o “Candy” ya que posiblemente se trate de sitios de pornografía infantil. Hay que evitar a toda costa la etiqueta CP.

Por el lado bueno de la Internet Invisible, podemos encontrar guías y listados telefónicos, e-mail y todo tipo de directorios, incluyendo listas de profesionales de cualquier disciplina. También podemos encontrar la venta de productos a través de e-commerce, leyes, decretos, casi cualquier tipo de información legal (aunque ésta también puede ser encontrada en webs estáticas), archivos multimedia y publicaciones digitales de libros y diarios. En esta parte también podemos encontrar sitios donde se comparten distintos conocimientos sobre sistemas, seguridad y muchas más cosas que sin duda resultan interesantes y no tienen ningún tipo de consecuencias para el usuario promedio.

Por el lado malo, en esta parte de Internet también podemos encontrar fácilmente, como ya hemos mencionado, pedofilia, venta de drogas, hackers, sicarios, películas hardcandy…También podemos encontrar manuales para fabricar bombas, venta de órganos, procedimientos para envenenar, mutilaciones, manuales de guerrilla, lavado de dinero y un sinfín de cosas más en torno a este campo. En la Internet Invisible se comercializa con drogas, armas, ácido sulfúrico,

servicios de hacker, secretos de estado, etc. Además hay sitios para el intercambio de pedofilia y cualquier otro tipo de ilegalidad que ya se ha mencionado, de las que se recomienda mantenerse muy atento para evitar caer en ellas.

Por otro lado también se encuentran temas relativos a conspiraciones de extraterrestres, avistamiento de ovnis y muchas más cosas relacionadas con el mundo paranormal.

Esta parte de Internet se considera como un mundo aparte ya que, por ejemplo, también posee su propio estilo de pago online llamado Bitcoins. La gran mayoría de las transacciones que se realizan aquí se llevan a cabo mediante las bitcoins, con la que se puede comprar virtualmente cualquier cosa. Bitcoin recurre a una base de datos distribuida en varios nodos de una red P2P para registrar las transacciones y utiliza la criptografía para proveer de funciones de seguridad como que las bitcoins sólo puedan gastarse por su dueño y no se puedan usar más de una vez. El diseño que presentan, permite poseer y transferir valor anónimamente. Además, pueden ser enviadas a través de Internet a cualquier usuario que tenga una “dirección Bitcoin”. (1 Bitcoin > $14 dólares).

Aquí podemos encontrar la Hidden Wiki, que es la enciclopedia del bajo mundo (una versión de la wikipedia en la Deep Web). En ella se puede encontrar un tutorial en el que nos explican cómo encontrar todo lo que buscas en este submundo de forma segura. Asimismo ofrece un listado de sitios que se pueden encontrar bajo el protocolo de Onionland.

También se pueden encontrar las páginas secretas de Google, que son páginas ocultas en las que aparecen imágenes extrañas de Google o con idiomas extravagantes como el klingon. Así como las páginas secretas de Mozilla Firefox en las que se pueden ver algunos comandos que sirven para obtener ciertos secretos y opciones del navegador.

Una parte de este submundo de la web está formado por las páginas que posiblemente reciban el mayor número de ataques por parte de los hackers y cuya información querría conocer más gente. Se trata de las páginas web de los gobiernos, FBI, CIA, Interpol o Bancos cuyo acceso suele realizarse únicamente por IP. Estas páginas son privadas y están ocultas por seguridad y son las más atacadas y a las que posiblemente sea más difícil el acceso de todo Internet.

La Deep Web también tiene una parte de leyenda. Hay una teoría que relaciona un gran número de estas páginas con los Illuminati. Esta teoría afirma que existen páginas con símbolos extraños o Illminati en las que si pinchamos nos redireccionan a páginas ocultas que revelan información de todo tipo, incluso del futuro (éstas serían las páginas Illuminati). Existen foros, páginas y vídeos en los que se encuentra información de estas webs.

Algunas de las páginas que podemos encontrar en la Internet Invisible son las siguientes:

- Hidden Wiki: como ya hemos explicado se trata de la “wikipedia” de esta parte de la web.

- CebollaChan: un sitio como 4chan pero en la web profunda.

- The Tor Library: es la página de la librería de TOR, en ella se puede encontrar libros sobre el nuevo orden mundial, manuales de programación, libros de medicina, negocios, cocina, etc.

- Quema Tu Móvil: es una página que pretende convencer a los usuarios para que dejen de utilizar el teléfono móvil.

- Tor Mail: que es el tipo de mensajería que se utiliza en esta web.

- Torch: es uno de los buscadores que tiene la Deep Web. Su uso es algo complicado ya que, al contrario de lo que hace Google, Torch no te envía sugerencias de búsqueda, por lo que el usuario tiene que saber específicamente qué busca.

- Tor University: es una página que ofrece servicios a estudiantes universitarios. En ella los estudiantes pagan una determinada cantidad de bitcoins para que se les haga algún trabajo de investigación y cosas por el estilo.

- Silk Road: es una página famosa dentro de la Deep Web que comercializa con todo tipo de drogas.

- Beidenmut: página en la que se pueden encontrar libros de fantasía, ocultismo, ciencia ficción, psicología, etc.

- Demonic Biblie-Magnus: donde puedes encontrar una biblia demoníaca escrita por Magnus, anticristo discípulo de Satanás.

6.12.- Ventajas y Desventajas

La principal ventaja que presenta la Internet Invisible es la gran cantidad de información que se puede encontrar en ella y la calidad de la misma ya que, teniendo en cuenta que los motores de búsqueda no tienen por qué ser calificadores de la calidad de la información, es más fácil encontrar documentos de calidad en una base de datos de 100.000 documentos que en una de 10.000. Por esta razón se conseguirán mejores resultados y más originales a la hora de realizar investigaciones en esta parte de la web. Asimismo, también podemos destacar como ventajas el anonimato, la privacidad y las facilidades que nos puede dar ante situaciones de opresión y coartación de la libertad de expresión.

En cuanto a las desventajas, podemos señalar la dificultad de acceso que existe a esta parte de la web sobre todo para los usuarios inexpertos ya que, la gran cantidad de información y los mecanismos que hay que seguir para poder acceder a ella pueden resultar abrumadores e incómodos. También cabe añadir el

grado de peligrosidad al que los usuarios se ven expuestos por acceder a la Internet Invisible, ya que no hay que olvidar que ésta no está controlada por estándares ni por organizaciones de seguridad informática.

Por otro lado, el anonimato también puede ser considerado una desventaja en los casos de los usuarios que acceden a la Deep Web sólo para realizar acciones ilegales (pedofilia, venta de armas, sicarios…), ya que les permite obrar con “total” libertad y complica la identificación de los mismos.

7.- Google e Internet Invisible

Como se ha comentado en apartados anteriores, hay algunos buscadores que están implementando o ya han implementado técnicas que permitan recuperar documentos pertenecientes a la Internet Invisible. En este apartado nos centraremos en algunos de los avances que ha realizado Google (por ser el motor de búsqueda actual más importante) para poder indexar páginas de esta parte de Internet.

7.1.- Crawlin a través de formularios HTML

En abril de 2008 se publicó un artículo en el Webmaster Central Blog de Google [18], en el que se afirmaba que Google estaba aumentando su interés por la Internet Invisible y estaba comenzando a aplicar técnicas de crawling para acceder e indexar los contenidos pertenecientes a esta parte de la web que resultasen interesantes.

Como se mencionó en apartados anteriores, un crawler (o robot) es un software que navega por la Web de forma automática y metódica con el objetivo de rastrear e indizar las páginas de la misma.

Resultaría muy interesante poder acceder a la Internet Invisible para extraer de forma automática sus contenidos. Para que este acceso sea posible, es necesario realizar consultas a las aplicaciones web, las cuales generan páginas dinámicas con la información relevante. Las aplicaciones web de la Internet Invisible únicamente proporcionan interfaces con formularios de búsqueda, en las que el usuario tiene que rellenar los campos para conseguir acceder a la información. Estos formularios, están pensados para ser utilizados por usuarios, por lo que, en un principio, no resultan procesables por un ordenador. Esta es una de las razones por la que se quería investigar el acceso a esta información de forma automática.

Los motores de búsqueda no son capaces de acceder a la Internet Invisible porque no pueden rellenar los formularios de búsqueda. Los contenidos de este tipo de aplicaciones quedan ocultos a los motores generales, por lo que deben buscarse otras vías para obtener estos contenidos.

Ciertos motores de búsqueda como Google, están avanzando en este campo y son capaces de realizar consultas a los formularios de aplicaciones de la Internet Invisible e indexar los contenidos. No obstante, las páginas indexadas pueden quedar obsoletas tanto por la información que incluyen como por cambios en los enlaces estáticos.

El equipo de Google ha investigado algunos formularios HTML con el fin de descubrir nuevas páginas web y URL que no podrían haber indexado de otra forma para las consultas que realizan los usuarios. El mecanismo que utilizan es, al encontrar una etiqueta <FORM> de un determinado sitio con calidad, realizan un

pequeño número de consultas a través de dicho formulario. Para los cuadros de texto, los ordenadores del equipo de Google, automáticamente eligen palabras del sitio que tiene el formulario. Por otra parte, cuando se encuentran con menús de selección, casillas de verificación y botones de opción en el formulario, se elige entre algunos de los valores de HTML. Una vez elegidos los valores para cada campo de entrada, generan y después intentan rastrear las URLs correspondientes a una consulta que un usuario haya hecho. Si la página resultante de la consulta presenta contenido interesante, válido y no se encuentra ya en su índice, se incorpora al índice como se haría con cualquier otra página. Esta práctica no se realiza en todas las páginas de la web, sino en un número limitado de sitios particularmente útiles.

Googlebot (el robot de Google) siempre respeta los ficheros robots.txt y las directivas noindex, nofollow, por lo que si algún formulario de búsqueda está prohibido en ese fichero, Googlebot no va a rastrear ninguna de las URLs que el formulario podría generar. Del mismo modo sólo recupera los formularios GET y evita los formularios que incluyen cualquier tipo de información del usuario. Se omiten todos los formularios que tengan un campo de contraseña o que usen términos asociados comúnmente con información personal como pueden ser logins, ids de usuarios, contactos, etc. Las páginas que se descubren con este tipo de rastreo no van en detrimento de las páginas que ya han sido rastreadas, por lo que este cambio no reduce el PageRank de las demás páginas. Este cambio tampoco afecta al rastreo, clasificación o selección de otras páginas web.

Esta investigación realizada por Google tiene el fin de aumentar su cobertura de la web ya que, desde hace tiempo, se pensaba que los formularios HTML eran la puerta de acceso a grandes volúmenes de datos inalcanzables por los motores de búsqueda. Con el rastreo de formularios HTML ha sido posible llevar a los usuarios de motores de búsqueda a encontrar documentos que de otra forma no se encuentran fácilmente y proporcionar a los usuarios una experiencia de búsqueda mejor y más completa.

Una forma de comprobar este rastreo de formularios, es realizando algunas consultas en Google, como por ejemplo buscar el libro de “El Hipnotista” (hacemos la búsqueda en inglés “the hypnotist book”) y el primer resultado que obtenemos presenta este libro en Amazon. Este es un claro ejemplo de acceso a la Internet Invisible desde un motor de búsqueda general ya que no se accede a la información rellenado el formulario de búsqueda de Amazon (Book Tittle: The Hypnotist) sino que se accede directamente a la página indexada en el buscador.

7.2.- Archivos PDF indexados

Google comenzó a indexar archivos PDF en 2001 y, actualmente cuenta con cientos de millones de archivos indexados en la web [19]. A diferencia de los documentos de texto estándar, los documentos escaneados (pdfs) no contienen datos de texto que puedan ser indexados por los robots de Google. En vez de eso,

Google utiliza un procedimiento llamado Reconocimiento óptico de caracteres (también llamado OCR), tecnología que transforma las palabras de fotos digitales en texto plano.

Esta nueva técnica permite que los documentos PDF se muestren directamente enlazados en el buscador, sin empeorar la calidad de las búsquedas. Hace mucho tiempo desde que Google muestra PDFs en sus resultados, pero antes basaba estas búsquedas en metadatos cercanos al documento. Sin embargo, ahora se ofrece la opción de ver el documento tanto en PDF como en HTML.

Google es capaz de indexar texto (en cualquier idioma) de los archivos PDF que utilizan distintos tipos de codificación de caracteres, siempre y cuando éstos no se encuentren cifrados ni protegidos por contraseña. Si el texto está insertado como imágenes, es posible procesarlas con algoritmos OCR para extraer el texto. Además, los enlaces que se encuentran incluidos en los archivos PDF son tratados de la misma forma que los que se encuentran en los documentos HTML, pueden indexarse, entrar en la clasificación PageRank y podemos seguirlos una vez rastreado el archivo PDF. Actualmente no se puede usar el atributo “nofollow” en los enlaces de un PDF.

7.3.- Archivos Flash indexados

En 2008, Google desarrolló un nuevo algoritmo capaz de indexar contenido textual en archivos Flash de todo tipo, ya fueran menús Flash, botones y banners para sitios web. Además, después se mejoró el funcionamiento del algoritmo por medio de la integración de la tecnología de Adobe Flash Player [20].

Este algoritmo explora los archivos Flash de la misma manera en que lo haría una persona: haciendo clic en los botones, introduciendo contenido y demás. El algoritmo recuerda todo el texto que encuentra y que luego estará disponible para ser indexado. La efectividad de este algoritmo fue mejorada con la utilización de la librería para búsqueda SWF de Adobe.

El motor de búsqueda de Google ya es capaz de indexar archivos en Flash, así como texto contenido en HTML, XML o el mismo Flash. También vincula el archivo Flash indexado al contenido descargado externamente y a los documentos de donde viene. Estas nuevas capacidades mejoran la capacidad de búsqueda al permitir que el contenido relevante en recursos externos aparezca en respuesta a las consultas de los usuarios [21].

Las palabras presentes en los archivos Flash se pueden usar para comparar con los términos de la consulta en las búsquedas de Google. Además, el texto de los archivos Flash puede utilizarse en las descripciones (snippets) de tu sitio web que muestra Google en los resultados de búsqueda. Asimismo, actualmente sólo se indexa el contenido de los textos de los archivos Flash, por lo que si estos archivos únicamente incluyen imágenes, no se van a reconocer ni indexar cualquier texto que pueda aparecer en esas imágenes.

Al igual que pasaba en el caso de los archivos PDF, en este tipo de archivos también se pueden rastrear las URLs que aparezcan en él. Estas mejoras realizadas no necesitan la intervención del propietario o diseñador de la web para que los contenidos puedan ser indexados.

8.- Conclusiones

Tras la realización de este proyecto se pueden sacar varias conclusiones atendiendo a los distintos temas que se han abordado en el mismo. La web es un dominio muy variable y complejo. Debido a su descentralización y a la aportación continua de elementos por parte de un gran número de usuarios, se encuentra en constante cambio, por lo que es muy difícil realizar análisis sobre la misma.

En primer lugar se citarán las conclusiones referidas al apartado de los buscadores. Actualmente, a pesar del gran número de tipos de buscadores que podemos encontrar, los más utilizados por los usuarios de la web son los motores de búsqueda, esto se debe a su facilidad de uso y su rapidez, lo que les hace especialmente atractivos a los usuarios. Cuando un usuario quiere buscar contenido en la web, en un motor de búsqueda simplemente tendrá que meter los términos de la consulta que quiera realizar en el campo de búsqueda.

Sin embargo, cuando el usuario quiere hacer consultas más generales se obtienen peores resultados que con el uso de un directorio, ya que almacena en la misma categoría todos los documentos referentes a un tema concreto, por lo que el usuario tendrá más información que consultar. Los directorios hoy en día, son utilizados en pequeñas comunidades cibernéticas para usos muy específicos. Esto se puede ver en Yahoo! que es uno de los directorios más importantes y actualmente sólo se usa para realizar búsquedas por palabras clave en la mayoría de los casos.

Como se vio en el apartado 3.7, los buscadores tienen una gran importancia en el uso actual de Internet, ya que se realizan más de 500 millones de búsquedas al día y constituye el segundo uso principal que se le da al Internet después del correo electrónico. Por esta razón, continúan investigando y buscando nuevas técnicas tanto para facilitar las búsquedas a los usuarios como para ser capaces de indexar y recolectar nuevas páginas en determinados formatos que antes no eran accesibles desde ellos.

El segundo apartado descrito en este documento es el relacionado con los ficheros robots.txt. Como conclusión a este apartado simplemente añadir que debería perfeccionarse el uso de estos ficheros ya que, como se ha podido ver, muchas veces los administradores de los sitios web exponen en los mismos una valiosa información que puede servir de gran ayuda a los hackers para realizar algún ataque. Por lo tanto, un fichero robots.txt tiene que estar creado por una persona cualificada y con los conocimientos de seguridad básicos.

En el tercer apartado se habla del proyecto Tor. Aunque este proyecto fue creado en 2003 con el fin de asegurar el anonimato al navegar por la red y mejorar la privacidad de los usuarios, hoy en día no es suficientemente conocido. Principalmente lo utilizan las personas relacionadas con el mundo de la informática o las que están interesadas en ella, pero la mayoría de los usuarios ni siquiera sabe de su existencia. Personalmente yo he sabido de este proyecto este

mismo año, cuando en una clase se habló por encima de él y se comentaron las oportunidades que ofrecía.

El penúltimo apartado está dedicado a esa parte de Internet no indexada por los buscadores, la Internet Invisible. En los últimos años se ha producido un gran avance en la variedad de formatos que pueden indizar los motores de búsqueda actuales, por lo que cada vez se puede acceder a más contenido de esta parte de Internet. Aún así, la relación entre el tamaño de la parte visible y de la invisible sigue siendo bastante desigual ya que, a pesar de poder indexar más documentos, también van apareciendo cada día nuevas páginas que se quedan en esta parte de la web. Personalmente, una de las grandes desventajas que encuentro en esta web es la gran cantidad de contenido ilegal que se puede encontrar en ella y la facilidad que se les concede a los usuarios para realizar acciones ilegales debido al anonimato con el que se debe acceder a ella.

Por otra parte creo que se debería trabajar en mejorar los accesos y los mecanismos de las páginas web presentes en la Deep Web que contienen contenido de calidad y que puede ser utilizado para proyectos de investigación. Considero que si eres un usuario básico es bastante difícil moverse por este tipo de web y, normalmente no consigues encontrar nada útil. Además, la falta de agilidad en este campo puede llevarte a visitar páginas con cierto contenido a las que nunca hubieses querido acceder.

El último apartado del documento trata de las mejoras que está realizando Google con el paso de los años para ampliar su cobertura de búsqueda en la web, indexando nuevos formatos y mejorando los resultados de búsqueda de cara a los usuarios finales. Como conclusión de este apartado cabe decir que van por el buen camino ya que Google es actualmente el motor de búsqueda que contiene el mayor número de páginas indexadas. Sólo les queda seguir trabajando por la misma vía y encontrar nuevas técnicas para penetrar en la web profunda e indexar el contenido presente en ella.

9.- Definiciones

.onion: pseudo dominio de nivel superior genérico (similar en concepto a los primigenios terminados en .bitnet y .uucp) que indica una dirección IP anónima accesible por medio de la red Tor.

AES-128: esquema de cifrado por bloques adoptado como un estándar de cifrado por el gobierno de los Estados Unidos.

Bootstrapping: proceso donde un sistema simple activa otro sistema más complejo para servir al mismo propósito.

Computación Cuántica: es un paradigma de computación diferente al de la computación clásica. Utiliza qubits en vez de bits y da lugar a nuevas puertas lógicas que permiten crear nuevos algoritmos.

Cookie: fragmento de información que se almacena en el disco duro del visitante de una página web a través de su navegador, a petición del servidor de la página.

Frame: imagen particular dentro de una sucesión de imágenes que componen una animación. La continua sucesión de estos fotogramas producen a la vista la sensación de movimiento, fenómeno dado por las pequeñas diferencias que hay entre cada uno de ellos.

Illuminati: sociedad secreta fundada en 1776 por Adam Weishaupt en Baviera, Alemania. Su meta era la mejora y el perfeccionamiento del mundo en el sentido de libertad, igualdad y fraternidad y la mejora y el perfeccionamiento de sus miembros.

JavaScript: lenguaje de programación interpretado. Se define como orientado a objetos, basado en prototipos, imperativo, débilmente tipado y dinámico.

Newsgroup (grupo de noticias): medio de comunicación dentro del sistema Usenet en el que los usuarios leen y envían mensajes textuales a distintos tablones distribuidos entre servidores con la posibilidad de enviar y contestar a los mensajes.

PageRank: es una marca de Google que contiene una familia de algoritmos utilizados para asignar de forma numérica la relevancia de los documentos indexados por un motor de búsqueda.

Plugin: aplicación que se relaciona con otra para aportarle una función nueva y generalmente muy específica.

Privoxy: programa que funciona como proxy web, usado frecuentemente en combinación con Tor y Squid. Tiene capacidades avanzadas de filtrado para

proteger la privacidad, modificar el contenido de las páginas web, administrar cookies, controlar accesos y eliminar anuncios, banners, ventanas emergentes y otros elementos indeseados de Internet.

Snippet: es la descripción mostrada debajo del resultado de la búsqueda en la lista de resultados. Informa al usuario que es lo que puede esperar del sitio si sigue el enlace.

Socks: protocolo de Internet que permite a las aplicaciones Cliente-servidor usar de forma transparente los servicios de un cortafuego de red.

TCP/IP: conjunto de protocolos de red en los que se basa Internet y que permiten la transmisión de datos entre computadoras. Los dos protocolos más importantes que lo componen son: TCP, protocolo que crea conexiones entre programas a través de las cuales pueden mandarse flujo de datos, e IP protocolo no orientado a conexión usado tanto por el origen como por el destino para la comunicación de datos.

Transport Layer Socket (en español, Seguridad de la Capa de Transporte): protocolo criptográfico que proporciona conexiones seguras por una red.

Webmaster: persona responsable del mantenimiento o programación de un sitio web.

Web ring: es un círculo de sitios web amigos que tienen tópicos en común y que se unen para promocionarse unos a otros. De esta manera el visitante que no encuentre cierta información en un sitio web será derivado a un sitio amigo en el web ring.

10.- Acrónimos

AES: Advanced Encryption Standard.

ARPA: Advanced Research Projects Agency.

DoD: Department of Defense.

FTP: File Transfer Protocol.

OCR: Optical Character Recognition.

P2P: Peer-to-peer.

SMPT: Simple Mail Transfer Protocol.

SRE: Standard for Robots Exclusion.

SSH: Secure SHel.

TLD: Top Level Domain.

TLS: Transport Layer Socket.

TCP/IP: Transmission Control Protocol / Internet Protocol.

11.- Bibliografía

[1] Maurice de Kunder. [Consultado 4 de Junio de 2012] Disponible en: http://www.worldwidewebsize.com/

[2] Netcraft. [Consultado 4 de Junio de 2012] Disponible en: http://news.netcraft.com/archives/2012/

[3] Boole, George. *An Investigation of the Laws of Thought.* 1854.

[4] Robertson, Stephen E. y Spärck Jones, Karen. Relevance weighting of search terms. *Journal of the American Society for Information Science*, 27:129-146, 1976.

[5] Robertson, Stephen E. The probability ranking principle in IR. *Journal of the American Society for Information Science*, 33:294-304, 1977.

[6] Granka, Laura; Joachims, Thorsten; Gay, Geri. "Eye-tracking analysis of user behavior in WWW search". En: *Procs of the 27th Annual international ACM SIGIR Conference on Research and Development in information Retrieval* (Sheffield, United Kingdom, July 25 - 29, 2004). New York: ACM Press, 2004, pp. 478-479.

[7] Marcos, Mari-Carmen; González-Caro, Cristina. "Comportamiento de los usuarios en la página de resultados de los buscadores. Un estudio basado en eye *tracking*". *El profesional de la información*, 2010, julio-agosto, v. 19, n. 4, pp. 348-358. Disponible en: http://grupoweb.upf.es/WRG/dctos/marcos_gonzalez_2010.pdf

[8] Bergman, Michael K. *The Deep Web: Surfacing Hidden Value* [white paper en Internet]. South Dakota: BrightPlanet Corporation; 2001 [Consultado 22 Mayo 2012].

[9] Aguillo, Isidro. *Herramientas avanzadas para la búsqueda de información médica en el web*. Aten Primaria 2002; 29 (4): 246-253

[10] Osorio, Silvia. Noticias.com [página principal en Internet]. Barcelona: Noticias Online [Citado 22 Mayo 2012]. La cara invisible de Internet; [2 pantallas].

[11] Red de Carreras de Comunicación Social y Periodismo [página principal en Internet]. Buenos Aires: Asociación Civil Red de Carreras de Comunicación Social y Periodismo-Redcom [Citado 17 Mayo 2012]. Encontrar lo que se busca: sumergiéndose en las profundidades; [aprox. 5 pantallas]. Disponible en: http://www.redcom.org/tp/parcial/g03/g03_parcial_base.htm

[12] Salazar García, Idoia. La Red Profunda: lo que los buscadores convencionales no encuentran [conferencia en Internet]. En: *Cultura & Política @ CiberEspacio:*

Actas del Primer Congreso ONLINE del Observatorio para la CiberSociedad; 2002. [Citado 22 Mayo 2012]. [1 pantalla].

[13] Hackeada la Web del Ministerio de Vivienda. En: *20 minutos.* 29.08.2007. [Citado 20 Mayo 2012]. Disponible en: http://www.20minutos.es/noticia/269614/0/hackean/web/vivienda/

[14] Fornas Carrasco, Ricardo. *La cara oculta de Internet* [en línea]. "Hipertext.net", núm. 1, 2003. [Consultado 21 Mayo 2012]. Disponible en: http://www.upf.edu/hipertextnet/numero-1/internet.html

[15] Ellsworth, Jill and Ellsworth, Matthew V.(1995). *Marketing on the Internet: Multimedia Strategies for the World Wide Web*. New York: John Wiley & Sons [Citado 23 Marzo 2012].

[16] Codina, Lluís. *Internet invisible y web semántica: ¿el futuro de los sistemas de información en línea?* [En línea]. [Citado 23 Marzo 2012]. Disponible en: http://www.lluiscodina.com/articulos/websemantica.pdf

[17] Sherman, Chris and Price, Gary. *The invisible Web*. Searcher.2001; 8(9):62-74.

[18] Madhavan, Jayant y Halevy, Alon. Crawling through HTML forms. Official Google Webmaster Central Blog, 2008. [Consultado 3 Junio 2012] Disponible en: http://googlewebmastercentral.blogspot.com/2008/04/crawling-through-html-forms.html

[19] Illyes, Gary. PDFs in Google search results. Official Google Webmaster Central Blog, 2011. [Consultado 3 de junio de 2012] Disponible en: http://googlewebmaster-es.blogspot.com.es/2011/09/archivos-pdf-en-los-resultados-de.html

[20] Adler, Ron, Stipins, Janis y Ohye, Maile. Improved Flash indexing. Google Webmaster Central Blog, 2008. [Consultado 3 de junio de 2012] Disponible en: http://googlewebmastercentral.blogspot.com.es/2008/06/improved-flash-indexing.html

[21] Stipins, Janis. Flash indexing with external resource loading. Google Webmaster Central Blog, 2009. [Consultado 3 de junio de 2012] Disponible en: http://googlewebmastercentral.blogspot.com.es/2009/06/flash-indexing-with-external-resource.html