Paakat
Revista de tecnología y sociedad

ISSN: 2007-3607

Universidad de Guadalajara
UDGVIRTUAL
suv.paakat@redudg.udg.mx
México

La Web Oculta y cómo los buscadores encuentran la información

José Antonio Amaro López
aljosea@hotmail.com
Universidad de Guadalajara, México

Ch. A. Lázaro
chlazaro@gmail.com
Universidad de Guadalajara, México

Gerardo Alberto Varela Navarro
gerardo@suv.udg.mx
Universidad de Guadalajara, México

Paakat: Revista de Tecnología y Sociedad, "Cultura digital y las nuevas formas del erotismo". Año 4, núm. 7, septiembre 2014-febrero 2015.

Recibido: 31/07/2014.
Aceptado para su publicación: 28/08/2014.

José Antonio Amaro López: licenciado en informática con orientación en sistemas computacionales; maestro en Tecnologías para el Aprendizaje, ambos por la Universidad de Guadalajara. Profesor docente del Departamento de Geografía y Ordenación Territorial del Centro Universitario de Ciencias Sociales y Humanidades de la Universidad de Guadalajara.
Lázaro Marcos Chávez Aceves: doctorado por El Colegio de Jalisco en Ciencias Sociales; docente de la Universidad de Guadalajara, UNIVA, entre otras instituciones educativas. Ha realizado investigaciones en temáticas de educación-tecnología, argumentación y diversidad sexual.
Gerardo Alberto Varela Navarro: licenciado en informática y Maestro en Tecnologías de la Información por la Universidad de Guadalajara (UDG), es académico e investigador del Sistema de Universidad Virtual de la UDG. Actualmente es doctorante del programa Sistemas y Ambientes Educativos de la UDGVirtual.

La Web Oculta y cómo los buscadores encuentran la información

José Antonio Amaro López

Ch. A. Lázaro

Gerardo Alberto Varela Navarro

Resumen

En el presente trabajo se aborda la diferencia entre lo que es la Web de la superficie y la Web Oculta, así como los problemas que en la actualidad deben resolver los buscadores para lograr indexar en sus bases de datos la mayor cantidad posible de sitios de la Internet, para proveer a sus usuarios de páginas que satisfagan sus necesidades de información.

Palabras clave

Web oculta, web de la superficie, rastreadores.

Deep Web and how search engines find information

Abstract

In this work, the difference between what the Web of the surface and the Hidden Web is discussed, and the problems that currently must meet to achieve the search engine in their databases as many of sites Internet, to provide their users pages that meet their information needs.

Keywords

Surface Web, d*eep web*, *crawlers*.

Introducción

Desde sus inicios la internet ha venido a revolucionar la manera en que se difunde la información, ya sea en sus inicios, mediante una simple página elaborada en el lenguaje HTML, o, en la actualidad, con las modernas páginas desarrolladas mediante combinaciones de diversos lenguajes de programación incrustados dentro del lenguaje HTML; lo anterior para incrementar la funcionalidad de la página y lograr una mejor interacción usuario-página.

Debido a la constante evolución de Internet y de la variedad de lenguajes de programación o de maquetación (JAVA, AJAX, HTML, CSS, HTML5, PHP) y manejadores de bases de datos (SQL, POSTGREL, MYSQL, entre otras), que pueden brindar la posibilidad de combinarse con el HTML, no sólo para satisfacer las necesidades de información o de ocio, sino para mejorar la interacción del usuario con la interne, ha complicado el sondeo de contenidos en la Internet por parte de los buscadores (Google, Yahoo, Bing, entre otros); lo que ocasiona que algunos sitios queden ocultos y no se contemplen dentro de los resultados que se arrojan en una búsqueda.

Desarrollo

Para hablar de la *Deep Web* primero debemos abordar lo que es la web de la superficie o *Surface Web*, la cual, según López-Barberá Martín , 2014, p. 96), es el conjunto de páginas que en la actualidad podemos consultar mediante los buscadores; es decir, todas aquellas páginas que no se encuentran elaboradoras en HTML[1], CSS[2], y que no contengan un for-

1 HyperText Markup Language o Lenguaje de marcas de hipertexto, es el lenguaje con el que se elaboran las páginas web.

2 Cascading Style Sheets u Hojas de Estilo en Cascada, se utiliza para maquetar las páginas elaboradas mediante el lenguaje HMTL.

mulario para acceder al contenido, además de que puedan ser indexadas por los buscadores mediante métodos que hacen uso del seguimiento de los enlaces que estas páginas contienen.

Ahora bien, la *Deep web* también llamada web oculta -la internet oculta- es la parte de internet a cuya información no es posible acceder de manera total mediante los buscadores, porque no es posible indexar las páginas de los sitios. Lo anterior debido a que el acceso, a las mismas, se encuentra restringido, ya sea por contraseña -como ocurre con los correos electrónico o sistema de bases de datos en línea de las empresas o de instituciones de gobierno- o mediante el llenado de un formulario que le permite al usuario solicita información para acceder a ésta; sirva de ejemplo, para este último caso, las bibliotecas virtuales y los formularios de páginas de comercio electrónico.

Algunos de Estos sitios ocultos contienen datos, artículos, estadísticas, libros, etc.; es decir, información valiosa de un amplio espectro para su utilización. En la actualidad el valor que tiene la información en la sociedad es alto, debido a la importancia que ésta tiene para la toma de decisiones; quien tiene acceso a ésta, de buena calidad o de fuentes primarias, además de lograr comprender el entorno donde se desenvuelve, podrá aprovechar mejor las oportunidades que se le presenten.

En años anteriores era relativamente fácil localizar la información en la Internet, ya que sólo se encontraba desarrollada en páginas que integraban los sitios mediante el lenguaje de etiquetado HTML, por lo que resultaba sencillo realizar la indexación del contenido por parte de los buscadores y sus rastreadores (*crawlers*). Pero con el auge de los *scripts* incrustados dentro del HTML y el uso de bases de datos, se ha complicado el rastreo e indexación de la información, dejándola oculta o imposible de localizar. Al encontrarse inaccesible la información para los usuarios y al considerarse de alto valor, Bergman realizó un estudio donde encontró que:

> La información pública en la Deep Web es de 400 a 500 veces más extensa que la contenida en la web de la superficie.
> La Deep Web contiene 7,500 terabytes de información en comparación con la web de la superficie que sólo contiene 99 terabytes.
> La Deep Web contiene cerca de 550 billones de documentos individuales comparado contra un billón que se localizan en la web de la superficie.
> Existían en el año 2000 más de 200,000 sitios web ocultos.
> Sesenta de los sitios más grandes de la Deep Web contenían cerca de 750 terabytes de información, lo que excedía 40 veces el tamaño de la información contenida por todos los sitios en la web de la superficie.
> En promedio, los sitios en la Deep Web reciben arriba del 50% del tráfico mensual que los sitios contenidos en la web de la superficie.
> La calidad de sus contenidos es de entre 1,000 a 2,000 veces más grande.
> El contenido de la Deep Web es altamente relevante para las necesidades actuales de información y de mercado.
> Aproximadamente el 95% de la información contenida en la Deep Web es información de acceso público, no sujeta a cuotas o suscripciones (Bergman 2001:1).

En otra investigación más reciente realizada con información del año 2004 por parte de He, *et. al*., (2004: 4-5), encontraron que la *Deep Web* creció del 2000 al 2004 entre 3-7 veces; es decir, creció en el orden de 10^5 sitios en estos años. Por estos motivos es necesario poder localizar esta información de mayor calidad y además abundante.

Como se mencionó con anterioridad, en la actualidad los buscadores hacen uso de rastreadores para que, de manera automática, busquen, clasifiquen, generen índices sobre el contenido y evalúan la calidad de la información de todos los sitios que se encuentran en la Internet para que las búsquedas sean más eficientes y contengan paginas relevantes que satisfagan las necesidades del usuario.

En un inicio era fácil indexar el contenido de la Internet debido a que se hacía un análisis de las etiquetas meta contenidas en las páginas HTML, donde se proporcionaba información acerca de la página y se analizaban las ligas contenidas en ellas para asignarles un nivel de importancia, para luego ser almacenadas por los buscadores. Posteriormente se consideró que un página contaba con un nivel de importancia alto, si a ésta hacían referencia otras páginas por parte de otros sitios de Internet; es decir, entre más recomendada era una página se considera como una fuente importante de información.

Lo anterior representa un esfuerzo relativamente fácil para lograr indexar las páginas de los sitios, ya que el acceso a éstas y sus contenidos eran públicos, siempre y cuando no estuviera el sitio localizado dentro de una *intranet*, además la página no se generaba de manera dinámica, pues una página HTML siempre existía en el servidor, aun cuando un usuario no la consultará. Sin embargo, en la actualidad existen muchos sitios de Internet donde las páginas que las integran se generan a petición del usuario; es decir, no existen hasta que se interactúa con el sitio, las páginas se generan en el momento por medio de los lenguajes de programación.

Como se puede ver, el problema surge cuando se trata de indexar las páginas que se generan de manera dinámica y además para acceder al contenido de éstas es necesario llenar formularios. También hay que agregar que para que el rastreador funcione, una vez que se le proporcionan una serie de direcciones URL de los sitios a revisar, debe contar con la capacidad de seguir las ligas que hacen referencia a otras páginas, las cuales también, en este tipo de páginas, se generan de manera dinámica las ligas o enlaces.

Ahora bien si las páginas que se generan son dinámicas, entonces ¿cómo sabrá el rastreador cuáles enlaces de las páginas seguir? Para resolver lo anterior Álvarez Díaz (2007: 32), en lista algunos problemas que debe resolver el rastreador con la finalidad de lograr indexar una página de la Web oculta y que algunos buscadores como Google ya resuelve:

- "Tratamiento del dinamismo del lado cliente (dinamismo de navegación)". Relacionado con la forma en la cual el usuario puede interactuar con la página, mediante el uso de menús generados de manera dinámica mediante algún lenguaje de programación, por lo cual es necesario que el rastreador pueda interpretar este tipo de navegación para poder seguir los diferentes enlaces de la página.

- "Descubrimiento de formularios". Una vez que el rastreador comprende cómo se encuentra estructurado el menú debe proceder a identificar qué tipo de página está revisando, si es HTML (página estática) o una generada de manera dinámica, para la primera no hay mayor problema, sólo revisar las etiquetas meta y revisar el contenido de la pagina para indexarla.

Sin embargo, para la página dinámica es necesario que el rastreador identifique si ha encontrado un formulario en la página y de qué tipo (si el formulario es del tipo buscador, o si es del tipo para ingreso a un sitio mediante nombre de usuario y contraseña o de consulta), si es HTML o generado por algún lenguaje de programación.

- "Modelado de formularios". Una vez que el rastreador identificó el tipo de formulario y el lenguaje que se utilizó para generarlo, se procede a modelarlo; es decir, el rastreador debe comprender cómo se encuentra estructurado el formulario para que se llenen los campos necesarios y así poder extraer la información del sitio de la web oculta.

- "Establecimiento de la relevancia del formulario para el *dominio de aplicación* y aprendizaje de la forma de realizar las consultas deseadas sobre el mismo". El rastreador deberá clasificar la página que se encuentra analizando y asignarle que tan relevante es el contenido que se presenta; es decir, decide si la información contenida se debe tomar en cuenta para indexarla, además deberá aprender a llenar el formulario o formularios que encuentre en otras páginas.

- "Generación de consultas sobre el formulario e invocación del mismo para obtener páginas de resultados". El rastreador una vez que comprende cómo se encuentra estructurado el formulario, procede a llenar de manera automatizada el formulario con base en lo aprendido por éste en el transcurso de su proceso de indexación de páginas. Entre más experto sea el rastreador, mejores búsquedas podrá realizar y por lo tanto podrá obtener páginas con contenido más relevante.

- "Estructuración de las páginas de resultados". Ya que se generan las consultas, el rastreador debe tener la capacidad de poder identificar la estructura que tiene la página para con esto analizar la información contenida en la misma y así decidir qué información se indexará.

- "Indexación y búsqueda, tanto de páginas de resultados como de los registros extraídos de las mismas". Para finalizar se debe diseñar un modelo que permita la indexación de la información que obtiene el rastreador y, posteriormente, se pueda tener acceso a los datos indexado y así proveer de resultados relevantes a los usuarios cada que estos ingresen una búsqueda.

Estos problemas ya han sido abordados por múltiples investigadores, lo que ha permitido generar distintas soluciones que presentan ciertas ventajas o desventajas al momento de indexar los sitios ocultos; ejemplo de lo anterior son el trabajo realizado por Supriya y Meenakshi, (2013), del Department of Computer Science, en India. Donde se propone un método para extraer datos de las páginas de la *Deep Web* utilizando el enfoque de construcción de consultas; se inicia con el ingreso de varias URL al modelo, éstas serán clasificadas por un módulo llamado *Web Classifier* el cual separa las páginas de la Web de la superficie de las que forman parte de la *Deep Web*. Una vez identificadas las páginas de la *Deep Web* se almacenan, se extraen las interfaces que contengan y se clasifican para, posteriormente, generar una interfaz donde el usuario llena los campos; estos se envían al modulo llamado *Query Builder que* genera las búsquedas y regresa los resultados al usuario.

También el trabajo de Madhavan, *et. al.*, (2009), de la empresa Google, se plante un método para clasificar las páginas de la *Deep Web*; los autores presentan algunas puntos importantes que han encontrado en el proceso de la construcción de su rastreador y se esbozan los principales retos que ellos encuentran para la exploración y uso de los contenidos presentes en la *Deep Web*; entre muchos otros.

Conclusión

Como puede verse, el trabajo para indexar en los sitios en la *Deep Web* resulta más complicado desde su consulta hasta su almacenamiento, pero también se han realizado trabajos que permiten encontrar estos tipos de páginas por la importancia y valor que representa la información que contienen.

Pero aun así existe un vacío que consiste en las páginas contenidas en la *Deep Web* las cuales no es posible indexar, ya que se hace uso de una tecnología llamada TOR (*The Onion Routing*). Ésta es una red abierta que le permite a los usuarios defenderse contra el análisis de tráfico que realizan algunas instancias gubernamentales sobre la Internet, y es una forma de vigilancia que amenaza la libertad personal y la privacidad, confidencialidad en los negocios, como de las relaciones, y la seguridad del Estado (TOR: Home, 2014) y fue un proyecto que creó el laboratorio de investigación naval de los Estados Unidos de Norte América (TOR: Overview, 2014).

Por lo tanto la *Deep Web* se vuelve aún más oculta o profunda dependiendo de cómo evolucioné la tecnología que permita mejorar la interacción del usuario con la Internet, y de que los buscadores tengan la capacidad de encontrar e indexar estas páginas; o si la intención es de hacer uso de estos avances con la finalidad de ocultar y así desarrollar de manera privada actividades que no se deseen dar a conocer.

Referencias

Álvarez Díaz, M. (2007). *Arquitectura para* Crawling *dirigido de información contenida en la web oculta* (Doctorado). Universidad da Coruña, Coruña.

Bergman, M. K. (2001). The Deep Web: surfacing hidden value, en *BrightPlanet-Deep Content.* A partir de http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.12.363&rep=rep1&type=pdf

He, B., Patel, M., Zhang, Z., y Chen-Chuan Chang, K. (2004). *Accessing the Deep Web: A Survey.* University of Illinois at Urbana-Champaing. http://www.inf.ufsc.br/~ronaldo/deepWeb/querying/Chang-dwsurvey-cacm07.pdf

López-Barberá Martín , A. (Abril de 2014). 'Deep Web' o Internet profundo. *Seguritecnia, revista decana independiente de seguridad* . Madrid, España.

Madhavan , J., Afanasiev , L., Antova , L., y Halevy, A. (2009). Harnessing the Deep Web: Present and Future, en *CiteSeerX.* A partir de http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.145.915&rep=rep1&type=pdf

Supriya & Meenakshi Sharma. (2013). Deep Web data mining, en *dInternational Journal of IT, Engineering and Applied Sciences Research*, 2 (3).

TOR: Home. TOR project: Anonymity Online: https://www.torproject.org

TOR: Overview. TOR Project: Anonymity Online: https://www.torproject.org/about/overview.html.en

¿Cómo citar?
AMARO LÓPEZ, J. A., Chávez Aceves. L. M. y Varela Navarro, G. A. (septiembre 2014-febrero 2015). La Web Oculta y cómo los buscadores encuentran la información, en *Paakat: Revista de Tecnología y Sociedad*, 4 (7).