Lámpsakos | N°.13 | pp. 39-50 | enero-junio | 2015 | ISSN: 2145-4086 | Medellín–Colombia

CrawNet: Crawler de Recursos Multimedia para la Web Superficial y Oculta

CrawNet: Multimedia Crawler Resources for Both Surface and Hidden Web

Alicia Martínez-Rebollar, Ph.D
Departamento en Ciencias de la Computación, Centro Nacional de Investigación Científica y Tecnológica
Cuernavaca, Morelos, México
amartinez@cenidet.edu.mx

Fernando Pech-May, MSc
Departamento en Ciencias de la Computación, Centro Nacional de Investigación Científica y Tecnológica
Cuernavaca, Morelos, México
fpech@cenidet.edu.mx

Hugo Estrada-Esquivel, Ph.D
Gerencia de Desarrollo de Nuevos Productos y Servicios, Fondo de información y documentación para la Industria, INFOTEC
Ciudad de México DF, México
hugo.estrada@infotec.com.mx

Eduardo Pedroza-Landa, MSc
Departamento en Ciencias de la Computación, Centro Nacional de Investigación Científica y Tecnológica
Cuernavaca, Morelos, México
eduardopl11c@cenidet.edu.mx

(Recibido el 20-10-2014. Aprobado el 20-12-2014)

Resumen. La web es la fuente de información de mayor uso en el ámbito académico, científico e industrial. Su crecimiento explosivo ha generado billones de páginas con información, las cuales se categorizan como web superficial, integrada por páginas estáticas que pueden ser indexadas; y web oculta, accesibles a través de formularios de búsqueda. En este artículo, se presenta el desarrollo de un crawler que permite realizar búsquedas, consultas y análisis de información en la web superficial y oculta en dominios específicos de la web.

Palabras clave: Crawler; web oculta; web superficial.

Abstract. The web is the most used information source in both academic, scientific and industry forums. Its explosive growth has generated billions of pages with information which may be categorized as surface web, composed of static pages that are indexed into a hidden web, accessible through search templates. This paper presents the development of a crawler that allows searching, queries, and analysis of information in the surface web and hidden in specific domains of the web.

Keywords: Crawler; hidden web; surface web.

Citación de artículo, estilo IEEE:
A. Martínez-Rebollar, F. Pech-May, H. Estrada-Esquivel, E. Pedroza-Landa, "CrawNet: Crawler de Recursos Multimedia para la Web Superficial y Oculta", *Lámpsakos*, N° 13, pp. 39-50 , 2015.

1. INTRODUCCIÓN

La World Wide Web ha tenido un crecimiento explosivo que ha generado grandes cantidades de información disponible en la web; actualmente representa la fuente de información más utilizada en el sector académico, científico e industrial.

Los motores de búsqueda son el mecanismo más utilizado para consultar la información contenida en la web. Estos motores coleccionan, de forma masiva, páginas que adquieren con la ayuda de rastreadores web (crawler), que atraviesan la web siguiendo los hipervínculos y almacenan las páginas en una gran base de datos para posteriormente indexarlas. De esta manera, cuando un usuario realiza una petición a un motor de búsqueda, este consulta su base de datos en la cual se han almacenado páginas indexadas a través de un crawler específico [1].

Un estudio realizado por BrightPlanet [2] demuestra que la mayoría de los motores de búsqueda solo acceden a información localizada en la web superficial, compuesta por documentos estáticos accesibles a través de ligas de navegación. Sin embargo, existe una gran cantidad de información en la denominada web oculta, solamente accesibles a través de formularios de búsqueda. Hoy en día, se hace difícil el acceso a información de la web oculta debido a que está constituida por bases de datos, formularios, contenido script, etc. que requieren la intervención del usuario [3].

En este artículo se presenta el desarrollo de un crawler, denominado CrawNet, de recursos multimedia para la web superficial y oculta, que permite realizar búsquedas, consultas y análisis de información en dominios específicos por medio de patrones de expresiones regulares que buscan relaciones semánticas entre conceptos.

El documento está organizado de la siguiente manera: en la sección 2, se describe el marco teórico; en la sección 3, se presenta la arquitectura del crawler para la web superficial y oculta; la sección 4 presenta los detalles de implementación; la sección 5 expone los resultados de la evaluación del Crawler; finalmente, en la sección 6 se presentan las conclusiones y trabajos futuros de esta investigación.

2. MARCO TEÓRICO Y TRABAJOS RELACIONADOS

2.1. Web superficial y oculta

La web superficial está constituida por páginas estáticas y que pueden ser indexadas por los motores de búsqueda tradicionales [4]. Según Fernández & Pardo [5], casi el 85% de los usuarios navegan en esta web. Se estima que el tamaño de la web oculta supera por mucho a la web superficial [6], [7]. En el año 2000, la web superficial era de 7,500 TeraBytes mientras que 167 TeraBytes eran de web oculta [8], [5].

La web oculta, también denominada web profunda o invisible, es aquella que contiene información que no es posible tener acceso a ella mediante un motor de búsqueda tradicional, ya que son casi inexplorados o no indexados [9] debido a que hace referencia a un contenido detrás de formularios HTML, por lo que es necesario rellenar formularios para realizar una consulta específica, autentificación de usuario [10] o la existencia de sitios dinámicos que utilizan scripts [11].

La web oculta contiene información considerada de gran valor [12]. Existen autores que proponen y muestran datos de estimaciones estadísticas acerca del tamaño actual de dicha información [13]. Sherman [14] propone una clasificación de la web oculta basándose en su contenido; dicha clasificación contempla 4 tipos de contenido: la web opaca, la web privada, la web propietaria y la verdadera web invisible.

2.2. Crawlers

Los crawlers son pequeños programas utilizados por los motores de búsqueda para la localización, proceso de análisis y almacenamiento de la información. Se clasifican en dos tipos: crawlers de la web superficial y crawlers de la web oculta. A continuación se describe cada una de estas categorías.

a. Crawlers para la web superficial. Son programas capaces de procesar y analizar la web. Se dedican a recorrer la web superficial localizando páginas estáticas para posteriormente extraer texto HTML e indexarlas [15].

b. Crawlers para la web oculta. Están enfocados en indexar contenido de páginas dinámicas, lo cual no pueden realizar los motores de búsqueda tradicionales, ya que la mayoría trabajan sobre páginas estáticas; sin embargo, carecen de efectividad en el manejo de información debido al gran tamaño de la web oculta; además, el acceso a las bases de datos depende de los datos introducidos en un formulario [16].

2.3. Trabajos relacionados

Actualmente se han propuesto diferentes trabajos de investigación que tienen como objetivo acceder a la web oculta [17], [18], [19]. En este artículo nos concentramos en presentar aquellos proyectos que cuentan ya con una herramienta que permite automatizar el proceso de búsqueda de información.

HiWE [20] es un crawler enfocado en la búsqueda de información del dominio de los semiconductores; usa técnicas para la extracción del contenido, tales como el cálculo de la distancia visual y uso de medidas de similitud textual para comparar los campos y los atributos del dominio. Para la extracción de información ocupa la técnica LITE (Layout Based Information Extration Technique), técnica que utiliza la extracción semántica, descarta todo lo que no sea texto, tales como imágenes, estilos de fuente y estilos de hoja y, finalmente, ayuda a seleccionar el candidato más apto para asociar una etiqueta con el elemento.

Google's Deep-web Crawl [21] tiene como objetivo indexar la mayor cantidad posible de páginas a partir de millones de formularios HTML. Permite acceder a la web oculta utilizando algoritmos para seleccionar valores de entrada para formularios que aceptan palabras clave y algoritmos para identificar entradas de formularios que solo aceptan valores de un tipo específico. Esto se realiza identificando los elementos HTML existentes en los formularios web (cajas de texto y menús de selección) e ignorando aquellos formularios que contienen información personal y áreas de texto para comentarios de usuarios. Su principal desventaja es que para identificar y enviar datos a formularios, utiliza plantillas de consulta que únicamente toman en consideración aquellos formularios con el método get.

DeepBot [3] es un buscador de la web oculta que recibe como entrada un conjunto de definiciones de dominio; utiliza diferentes heurísticas para identificar automáticamente formularios de consulta relevantes. DeepBot emplea mini buscadores web automatizados; utiliza algoritmos heurísticos que miden la distancia entre los elementos y las etiquetas, de tal forma que puedan asociarse semánticamente los campos existentes. El esquema propuesta contiene los elementos del dominio y las relaciones entre ellos y determina los formularios relevantes y la forma en que los campos del formulario deben ser completados.

3. ARQUITECTURA DEL CRAWLER PARA LA WEB SUPERFICIAL Y OCULTA

La metodología propuesta busca solucionar la recuperación de información específica que se encuentra en la web superficial y oculta en un dominio específico. Para esto se desarrollaron dos algoritmos con el propósito de aplicar esta solución; los algoritmos que integran nuestra metodología buscan identificar y recuperar información solicitada. CrawNet recibe como entrada una semilla URL, utilizada como un dominio web, para realizar la búsqueda. En la Figura 1 puede apreciarse la metodología para el desarrollo de CrawNet para la web superficial y oculta, y a continuación se describe cada algoritmo.

3.1. Algoritmo para el acceso a la web superficial

El algoritmo permite la recuperación de información y enlaces (URLs) basado en cuatro tipos de búsquedas (general, videos de fútbol, gamecast e información de películas); para lograrlo, recibe la semilla URL para posteriormente recorrer y recuperar todos los enlaces existentes en el contenido HTML. Cada enlace recuperado es comparado y evaluado dependiendo del tipo de búsqueda seleccionado.

Para la evaluación y selección de URLs se declararon patrones de expresiones regulares que contienen palabras clave que se comparan directamente; los URLs que coinciden con los patrones se recuperan y se almacenan en una base de datos (BD). A continuación se describe cada etapa.

A. *Búsqueda de URLs con información solicitada por el usuario*. Busca e identifica URLs de acuerdo al tipo de búsqueda, dando como resultado la recuperación de enlaces candidatos que pueden contener la información solicitada. Cada enlace pasa por filtros definidos por patrones de expresiones regulares, integrados por palabras clave relacionadas semánti-

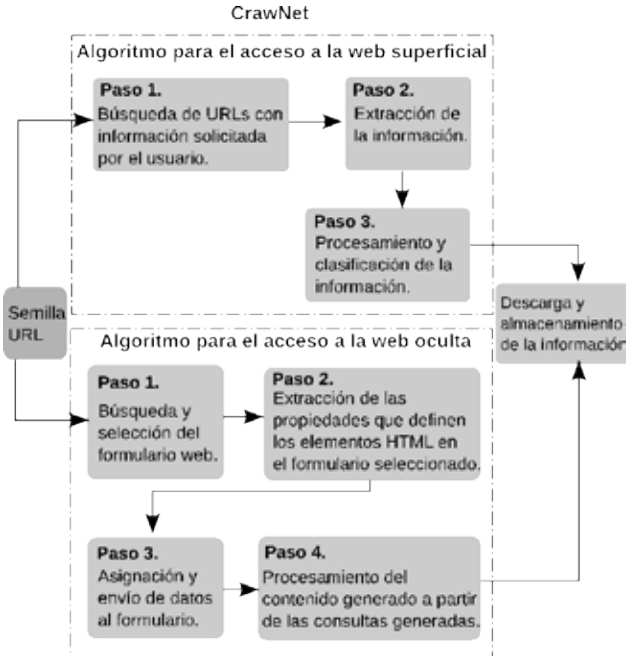

Fig. 1. Metodología CrawNet para la web superficial y oculta.

camente al tipo de búsqueda. Asimismo se encarga de buscar coincidencias entre las palabras clave que integran los patrones de expresiones regulares y las cadenas de texto que conforman un URL. Cuando las palabras clave del patrón coinciden con alguna palabra que se encuentre dentro de un enlace, se selecciona para ser procesado en el siguiente paso.

B. *Extracción de la información.* Se recibe y procesa el URL seleccionado, el proceso identifica y extrae el enlace si este contiene la información solicitada, basado en los cuatro tipos de búsqueda.

C. *Procesamiento y clasificación de la información.* Se procesa la información extraída en la etapa anterior. El proceso clasifica y almacena los URLs en una BD. Los URLs y la información recuperada se procesan de acuerdo al tipo de búsqueda utilizada.

3.2. Algoritmo para el acceso a la web oculta

Permite identificar páginas web que contienen formularios de búsqueda avanzada y recursos multimedia. Dichos formularios solicitan campos o detalles específicos de un dominio en particular. Los formularios de búsqueda avanzada utilizan un mayor número de campos para describir los elementos buscados y permite que las respuestas sean más cercanas a la solicitada por el usuario. Se tomaron como formularios de búsqueda avanzada aquellos que solicitan

Tabla 1. Patrón para la búsqueda de enlaces y formularios de búsqueda avanzada.

Nombre de patrón	Palabra clave
form_ avanzado	Search, advanced-search, advanced_search, advancedsearch, search_advanced, searchadvanced, advanced, adv_search|adv-search, advsearch, searchadv, buscar, busquedaavanzada, busqueda_avanzada, busqueda-avanzada, avanzada-busqueda, avanzada_busqueda, busqueda avanzada, avanzada

como mínimo tres campos. CrawNet es capaz de enviar datos a través de dichos formularios para obtener una respuesta y analizarla. Para realizar esto, el algoritmo se basa en cuatro pasos principales.

A. *Búsqueda y selección del formulario web.* Realiza la búsqueda de parámetros denominados semillas (por ejemplo: título, actor, género y director para películas cinematográficas) y que se encuentran en los formularios de búsqueda avanzada. Para enviar los datos al formulario, es posible usar un solo parámetro, pero se recomienda el uso de más parámetros para la obtención de mejores resultados. La búsqueda en la web se lleva a cabo mediante la búsqueda por expresiones regulares (consiste en identificar y seleccionar un URL como candidato potencial) y por búsqueda de manera exhaustiva (busca directamente en cada web la etiqueta <form>).

B. *Extracción de las propiedades que definen los elementos HTML en el formulario seleccionado.* Se extraen, mediante un parser, los valores de los atributos de los elementos HTML (form, textbox, combobox y button) los cuales se relacionan semánticamente con los parámetros (título, actor, director y género). De esta manera, se identifica la relación entre las etiquetas y los campos a utilizar para ser llenados por la metodología CrawNet. La extracción de los valores de las propiedades (name, value, action, id) del formulario se realiza mediante el parser Jsoup, dichos valores deben coincidir con el patrón establecido de la Tabla 1.

C. *Asignación y envío de datos al formulario.* Se realiza el envío de datos al formulario web a través de los lenguajes de navegación HtmlUnit y Selenium, que simulan operaciones de un usuario común sobre páginas web. Estos lenguajes utilizan los valores de las propiedades de los elementos HTML para enviar los datos al formulario;

para lograrlo, se proporciona el valor del atributo del formulario, se le asigna valores semillas a los elementos textbox y/o combobox y finalmente se realiza la consulta con el botón.

D. *Procesamiento del contenido generado a partir de las consultas realizadas.* Se procesa la información obtenida a partir de la consulta realizada por el CrawNet mediante búsquedas de palabras que identifican la búsqueda solicitada dentro del contenido HTML (obtenido como respuesta). Si el elemento es encontrado, el URL es almacenado en la BD. Para analizar la respuesta generada, se lee por completo el documento HTML para aperturar los enlaces que contiene; cada enlace se procesa en busca de valores semillas.

CrawNet fue diseñado para la extracción de recursos multimedia en el dominio cinematográfico, pero adaptable a diferentes dominios. Su arquitectura se encuentra dividida en cinco módulos (ver Figura 2). Los dos primeros módulos se utilizan para ambas búsquedas (superficial y oculta). La entrada principal de esta arquitectura es una semilla URL que representa el dominio web donde ambas búsquedas realizan el proceso de recuperación de información. La búsqueda sobre la web oculta requiere datos adicionales llamados valores semilla (que se utilizan para identificar una semilla). A continuación, se describen cada uno de los módulos que se muestran en la Figura 2.

- *Recuperación de páginas.* Realiza la recuperación y almacenamiento de los URLs contenidos dentro del dominio especificado en la semilla. Como primer paso identifica, obtiene y almacena todos los URLs encontrados en el dominio para posteriormente almacenarlo en una BD. Posteriormente se accede a la BD para recuperar el primer URL e inspeccionarlo a profundidad para la obtención de nuevos enlaces; cada URL es navegado y almacenado en una BD. La búsqueda es finalizada cuando no existe un nuevo URL dentro de las páginas almacenadas.

- *Analizador y clasificador de páginas.* Se encarga de procesar cada enlace seleccionado de acuerdo al tipo de búsqueda. Para la web superficial se utilizan filtros sobre los URLs que identifican y envían los enlaces a las clases respectivas para su procesamiento; posteriormente se almacena en la BD. Para la web oculta se identifican páginas web candidatas

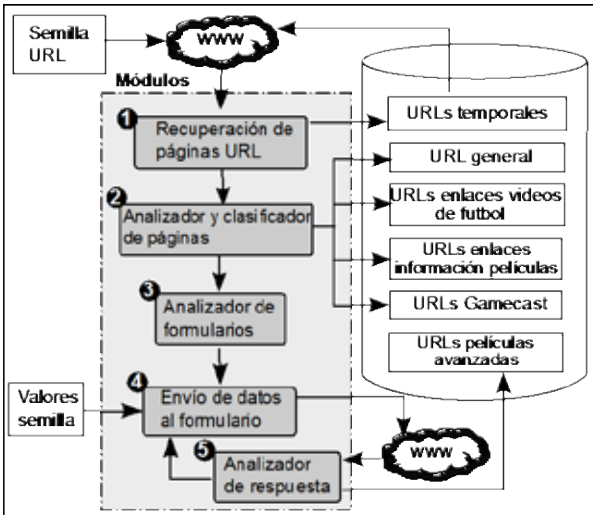

Fig. 2. Panorama general de la arquitectura CrawNet.

a contener formularios de búsqueda avanzada. Esta identificación de páginas candidatas se realiza paralelamente con el primer módulo de dos diferentes maneras: En la primera, se realiza la búsqueda de palabras clave que den indicios de la presencia de un formulario de búsqueda avanzada; para este caso se utiliza la palabra clave "búsqueda avanzada" junto con sus derivaciones y traducciones (search, advanced-search, etc.), las cuales son basadas en las propiedades name, action e id de los formularios; con la segunda manera, se identifican las etiquetas <form> dentro del contenido de las páginas.

- *Analizador de formularios.* Identifica y recupera formularios de búsqueda avanzada junto con los elementos del <form> (textbox, combobox y button) relacionados a un dominio específico; esto se realiza buscando similitud entre los parámetros de dominio, etiquetas y propiedades (name, value e id) de los elementos HTML identificados como candidatos. Se usaron como parámetros de dominio las palabras título, actor, director y género (así como sus derivaciones y traducciones) relacionadas con el tópico de películas cinematográficas. Un formulario de búsqueda es seleccionado solo en caso de haber encontrado al menos tres correspondencias entre elementos del formulario y parámetros de dominio. Finalmente, se identifican valores semilla que servirán para realizar una consulta a través del formulario de búsqueda, obteniendo los valores apropiados para los elementos HTML identificados.

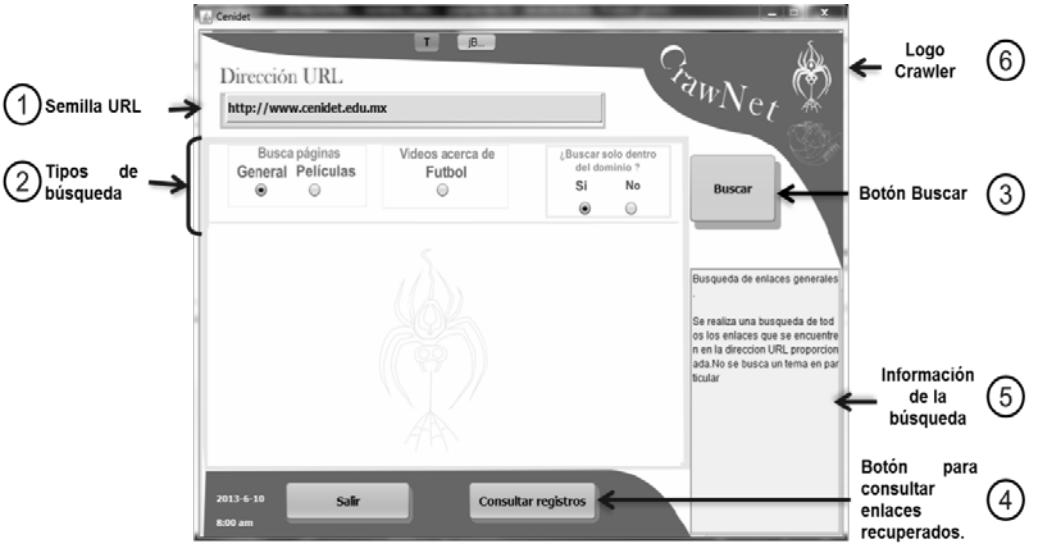

Fig. 3. Interfaz principal de la herramienta CrawNet.

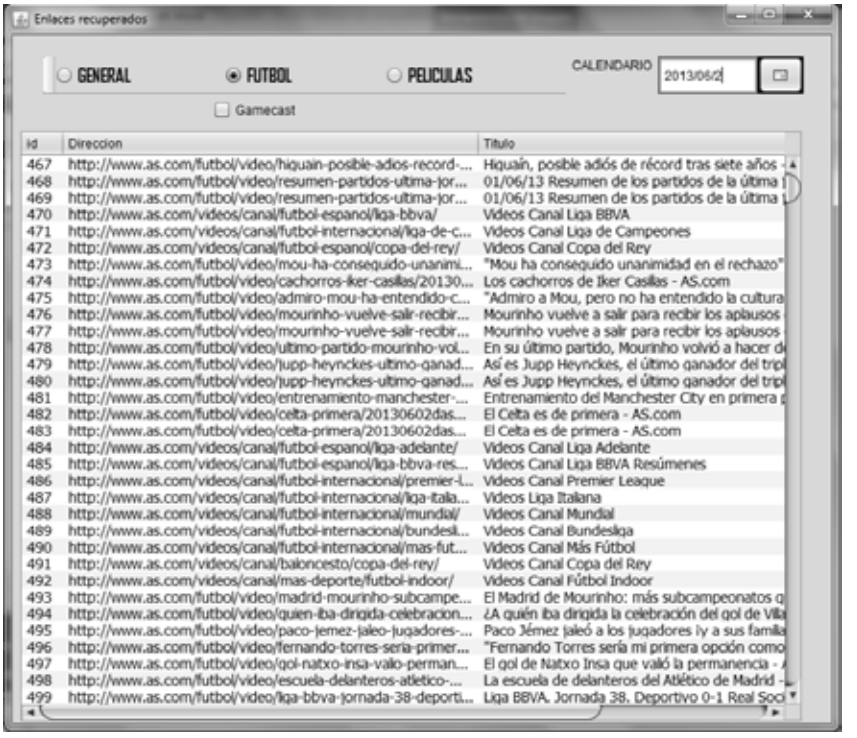

Fig. 4. Interfaz de enlaces recuperados.

- *Envío de datos al formulario.* Asigna y envía los valores semilla al formulario de búsqueda seleccionado para realizar la consulta. Los valores semilla son asignados al elemento HTML correspondiente del formulario analizado. Posteriormente, se realiza la simulación del evento clic en el elemento button mediante HtmlUnit y Selenium.

- *Analizador de respuesta.* Evalúa la respuesta obtenida a partir de la consulta realizada al formulario de búsqueda a través de un análisis completo del contenido HTML de la página resultante. Cabe mencionar que la mayoría de los buscadores presentan los resultados de búsqueda en un mecanismo de paginación; para este tipo de buscadores, solamente se analiza el contenido de la primera página de resultados y tomando en consideración los resultados de mayor relevancia encontrados en la página. La respuesta apropiada se almacena en la BD.

El resultado final del proceso realizado por esta arquitectura es reflejado en los registros de una tabla, la cual contiene un conjunto de enlaces con información relacionada con las palabras semilla.

4. IMPLEMENTACIÓN

La Figura 3 muestra la interfaz principal de CrawNet desarrollado bajo la plataforma Java. La interfaz gráfica permite al usuario recuperar información de manera asistida. Los enlaces resultantes pueden ser visualizados sin necesidad de concluir el proceso de búsqueda de información.

Como puede apreciarse en la interfaz, la semilla URL es agregada en un campo de texto; el usuario podrá seleccionar diferentes tipos de búsqueda, tales como general, películas, videos, gamecast y películas avanzadas. Después que el usuario seleccione sus opciones de búsqueda, el usuario podrá ejecutar la consulta.

El resultado de la búsqueda es una lista de enlaces recuperados de una consulta en un dominio específico. En la Figura 4 puede apreciarse un ejemplo de los resultados de enlaces recuperados en una búsqueda de videos de fútbol.

Fig. 5. Implementación de búsqueda avanzada.

La implementación de la herramienta CrawNet fue realizada en sitios reales de internet. En el campo de semilla URL se agrega una dirección URL relacionado con cualquiera de cuatro tipos de búsqueda y, en cualquiera de los casos, se retorna una lista de enlaces recuperados (Figura 4). Para el caso de búsqueda en la web oculta, se realiza a través de la búsqueda avanzada donde se requiere rellenar campos para la búsqueda (Figura 5).

5. EVALUACIÓN DE RESULTADOS

Para la evaluación del crawler se aplicaron las métricas de precisión, exhaustividad y medida F [22]. Estas métricas son utilizadas en la evaluación de los sistemas de recuperación de información (SRI). La precisión determina el porcentaje de aciertos de una operación en un SRI [23]. La exhaustividad determina el porcentaje de datos relevantes recuperados respecto al total de datos relevantes existentes en una BD [23]. La medida F (o medida armónica f) determina el promedio existente entre los valores de precisión y la exhaustividad [22].

Como se ha mencionado anteriormente, el algoritmo de la web superficial permite búsquedas generales, videos de fútbol, información de películas y gamecast. En el caso de búsquedas generales, no se aplicaron la precisión y exhaustividad por la carencia de filtros; sin embargo, se analizó el tiempo de consulta y el número de enlaces recuperados dentro y fuera del dominio (ver Tabla 2). En la Tabla 2 solo se muestra el resultado de los primeros 5 de los 120 enlaces evaluados.

Tabla 2. Resultados de caso de estudio de búsqueda general.

Sitio	Número de enlaces recuperados			
	En el dominio		Fuera del dominio	
	1 hora	2 horas	1 hora	2 horas
http://msn.foxsports.com	8178	10622	9033	18170
www.sport.es	2633	3380	8985	11481
www.cnn.com	2634	4297	12720	18720
www.amazon.com	10646	17483	22375	30837
www.youtube.com	18089	21443	14240	24170

En la Tabla 3 se puede apreciar los dominios seleccionados para los otros tipos de búsqueda. Se seleccionaron grupos de 100 enlaces que se utilizaron como muestra representativa de cada dominio; cada enlace fue almacenado en la BD para posteriormente realizar el análisis e Identificación de Elementos Relevantes (IER) y aplicar la métrica de precisión. Asimismo, la tabla muestra la IER Realizada Manualmente (IERM) y la IER mediante CrawNet (IERC) sobre las 100 muestras; por último, se aprecia el porcentaje de la Precisión (P) de la búsqueda.

La evaluación de los casos de estudio de la web oculta se dividió en “selección de formularios y elementos” (que evalúa la correcta selección del formulario) y “evaluación de respuesta generada” (que evalúa las respuestas generadas de las consultas realizadas por CrawNet). La evaluación de las dos secciones implicó una verificación directa sobre los dominios web con el fin de identificar formularios de búsqueda avanzada, para así registrar los URLs, etiquetas, elementos, etc. Posteriormente, se realiza la consulta para verificar los URLs que se obtenían como resultado.

5.1. Selección de formularios y elementos

Las pruebas se realizaron en 18 sitios (ver columna 1 de Tabla 4). La Tabla 4 también muestra un resumen de los resultados obtenidos de los 18 sitios en la identificación y selección de los elementos HTML (selección y recuperación del formulario y asociación de parámetros con valor de atributos en textbox y combobox). Puede apreciarse que 16 sitios obtuvieron el 100% en Precisión (P), Exhaustividad (E) y Medida F (M); en 2 sitios no se pudieron identificar formularios de búsqueda avanzada.

Por cada sitio, se realizaron dos tipos de verificaciones, manualmente y utilizando CrawNet. En la Tabla 5 se muestran los elementos descriptivos que sirvieron para realizar la comparación entre los dos tipos de evaluaciones (columna 1). Debido a la falta de espacio, solo se muestra el análisis e identificación de parámetros del sitio www.filmsandtv.com; en dicho sitio, se seleccionó 1 de 2 formularios existentes en la página por lo que su P, E y M fue del 100%.

Tabla 3. Resultados de la búsqueda en dominios específicos. Identificación de Enlaces Relevantes Manualmente (IERM), Identificación de Enlaces Relevantes mediante CrawNet (IERC).

Tipos de búsqueda	Dominio URL	IERM	IERC	P
Videos de fútbol	www.as.com	96	96/100	96%
	www.fifa.com	92	92/100	92%
	http://espnfc.com	95	95/100	95%
	http://futbol.univision.com	67	67/100	67%
	www.marca.com	100	100/100	100%
	Promedio total: 90%			
Info. de películas	www.accionhd.com	94	94/100	94%
	www.allmovie.com	100	100/100	100%
	www.amazon.es	88	88/100	88%
	www.elmultinice.com	100	100/100	100%
	www.estrenosdecine.net	100	100/100	100%
	Promedio total: 96%			
Gamecast	http://espndeportes.espn.go.com	100	100/100	100%
	http://espnfc.com	100	100/100	100%
	Promedio total: 100%			

Los resultados obtenidos mediante la evaluación de las métricas de precisión y exhaustividad sobre la información recuperada arrojaron porcentajes muy altos, que, en la mayoría de los casos, fue de un promedio del 100%; sin embargo, este porcentaje se obtuvo sobre aquellos URLs que el algoritmo de la web oculta logró recuperar al identificar un formulario de búsqueda avanzada en su contenido. Los enlaces en los que no se llegó a identificar este tipo de formulario no pudieron ser evaluados. La razón por la que no se identificaron algunos formularios de búsqueda avanzada fue que existieron enlaces que no coincidieron con ningún patrón de búsqueda avanzada establecido.

Tabla 4. Resultados de la identificación y selección de elementos HTML. Selección y Recuperación del Formulario (SRF), Asociación de parámetros con valor de atributos en TextBox y Combobox (AP-TC) e Identificación y Selección de Botón (ISB).

Enlace URL	SRF			AP-TC			ISB		
	P	E	M	P	E	M	P	E	M
www.filmsandtv.com	100%	100%	100%	100%	100%	100%	100%	100%	100%
www.findanyfil.com	100%	100%	100%	100%	100%	100%	100%	100%	100%
www.half.ebay.com	100%	100%	100%	100%	100%	100%	100%	100%	100%
www.starscafe.com	100%	100%	100%	100%	100%	100%	100%	100%	100%
www.videomania.info/es	100%	100%	100%	100%	100%	100%	100%	100%	100%
www.bbfc.co.uk	100%	100%	100%	100%	100%	100%	100%	100%	100%
www.accionhd.com	100%	100%	100%	100%	100%	100%	100%	100%	100%
www.swapadvd.com	100%	100%	100%	100%	100%	100%	100%	100%	100%
www.chapter.indigo.ca	100%	100%	100%	100%	100%	100%	100%	100%	100%
www.dvdgo.com	100%	100%	100%	100%	100%	100%	100%	100%	100%
www.cinegratis.net	100%	100%	100%	100%	100%	100%	100%	100%	100%
www.k12digitalmovies.com	100%	100%	100%	100%	100%	100%	100%	100%	100%
www.movlic.com	100%	100%	100%	100%	100%	100%	100%	100%	100%
www.megamovieline.com	100%	100%	100%	100%	100%	100%	100%	100%	100%
www.movietickets.com	100%	100%	100%	100%	100%	100%	100%	100%	100%
www.hispashare.com	100%	100%	100%	100%	100%	100%	100%	100%	100%
www.amazon.es	0%	0%	0%	0%	0%	0%	0%	0%	0%
www.amazon.com	0%	0%	0%	0%	0%	0%	0%	0%	0%
Total	88.8%	88.8%	88.8%	88.8%	88.8%	88.8%	88.8%	88.8%	88.8%

Tabla 5. Análisis e identificación de parámetros en formularios web del sitio www.filmsandtv.com.

	Valores encontrados	
Parámetros evaluados	Verificación manual	Verificación CrawNet
Enlace semilla	www.filmsandtv.com	www.filmsandtv.com
Enlace formulario	www.filmsandtv.com/advsearch.php	www.filmsandtv.com/advsearch.php
Formulario seleccionado	Name=”frmAdvSrch”	Name=”frmAdvSrch”
Número de etiquetas descriptivas relevantes	4	4
Dominio de etiquetas descriptivas relevantes	Title, actor, director, genre	Title, actor, director, genre
Número de elementos asignables relevantes	4	4
Dominio de elementos asignables	Título =”asTitle” Actor =”asActor1” Director =”asDirector1” Género = “asGenre”	Título = ”asTitle” Actor = ”asActor1” Director = ”asDirector1” Género = “asGenre”
Botón seleccionado	Name=”btnAdvSearch”	Name=”btnAdvSearch”

Tabla 6. Resultados de evaluación de una película en 8 sitios web distintos con 3 Tipos de Búsqueda (TB): 1) por Título (T), 2) por Título y Actor (TA) y 3) por Título, Actor y Director (TD).

Enlaces URL	TB	TER	TERu	TERr	P	E	M
www.filmsandtv.com	T	6	101	6	5.9%	100%	11.14%
	TA	4	51	4	7.8%	100%	14.47%
	TD	2	31	2	6.4%	100%	12.03%
www.findanyfilm.com	T	15	23	15	65.22%	100%	78.95%
	TA	5	5	5	100%	100%	100%
	TD	3	3	3	100%	100%	100%
www.half.ebay.com	T	18	20	18	90%	100%	94.74%
	TA	18	12	10	83.33%	55.5%	66.67%
	TD	14	8	8	100%	100%	100%
www.bbfc.com.uk	T	20	81	15	18.52%	75%	30.07%
	TA	29	29	29	100%	100%	100%
	TD	9	9	9	100%	100%	100%
www.accionhd.com	T	7	8	7	87.50%	100%	93.33%
	TA	3	3	3	100%	100%	100%
	TD	0	0	0	-	-	-
www.swapadvd.com	T	10	26	3	11.54%	30%	16.67%
	TA	10	21	6	28.57%	60%	38.71%
	TD	5	6	5	83.33%	100%	69.77%
www.videomania.info	T	2	2	2	100%	100%	100%
	TA	1	2	1	50%	100%	66.67%
	TD	0	0	0	-	-	-
www.starscafe.com	T	10	174	10	5.75%	100%	10.87%
	TA	10	166	10	6.02%	100%	11.36%
	TD	10	3	2	66.67%	20%	30.77%

5.2. Evaluación de respuesta generada

Las pruebas realizadas en esta sección se realizaron en 8 sitios web (ver columna 1 de Tabla 6), los cuales fueron procesados por CrawNet. Por cada sitio se realizaron consultas de diez películas cinematográficas de forma manual y usando el algoritmo de la web oculta; sin embargo, debido a la carencia de espacio, únicamente se muestra el resultado de la búsqueda de una película (Título: Superman, Actor: Christopher Reeve, Director: Richard Lester) en cada sitio (Tabla 6).

En la Tabla 5 se muestra el resultado de la consulta de dicha película en la web oculta. Las pruebas en cada sitio web involucraron 3 tipos de búsqueda: búsqueda por título, búsqueda por título y actor y búsqueda por título, actor y director. Como primer paso, se analizó manualmente (en cada sitio) el Total de Enlaces Relevantes (TER), los cuales corresponden a los enlaces de las películas en cada sitio. Asimismo se evaluó el Total de Enlaces Recuperados (TERu) por CrawNet y el Total de Enlaces Relevantes Recuperados (TERr) por CrawNet. Finalmente, se muestran los resultados P, E y F.

Los resultados obtenidos basados en la información recuperada arrojaron porcentajes muy altos en algunos casos. Los resultados sobre los ocho dominios web procesados siguieron un mismo comportamiento. CrawNet recuperó muchos más enlaces que los que debía recuperar, afectando con esto la precisión; sin embargo, se puede afirmar que se lograron recuperar todos los URLs esperados, dando como consecuencia porcentajes muy altos en cuanto a la exhaustividad del algoritmo.

6. CONCLUSIONES Y TRABAJOS FUTUROS

En las búsquedas permitidas en el algoritmo de la web superficial, se utilizaron patrones de expresiones regulares para poder filtrar la información recuperada. Se analizaron un total de 1200 URLs de forma manual para corroborar que la recuperación

fuera de acuerdo a la búsqueda seleccionada. Los resultados obtenidos de la web superficial fueron superiores al 90% en precisión; por lo tanto, se comprobó la efectividad del algoritmo desarrollado.

El algoritmo de la web oculta presentó retos durante su desarrollo ya que debía identificar páginas que tuvieran formularios de búsqueda avanzada, determinar el dominio de películas al que pertenecía y comprobar que el algoritmo lograra enviar datos al formulario para analizar la respuesta generada a dicha consulta.

El algoritmo de la web oculta logró identificar formularios de búsqueda avanzada en la mayoría de los casos; sin embargo, los resultados presentados fueron a partir de solo 18 sitios que permitían la búsqueda avanzada de películas cinematográficas. En los resultados se pueden apreciar que tuvo un 100% en las métricas de exhaustividad. Para estas pruebas puede parecer bastante alto, pero es importante aclarar que el porcentaje reportado es solo de los sitios en los que se pudo identificar adecuadamente el formulario de búsqueda avanzada; los demás sitios web en los que no se pudo identificar, no se tomaron en cuenta.

Actualmente se está trabajando en la generación de formularios web de manera automática, utilizando técnicas y metodologías actuales; asimismo, pretendemos utilizar ontologías para analizar, estructurar y organizar información de tal manera que permita descubrir patrones que siguen ciertos temas en particular para lograr la generación virtual de estos formularios y, una vez generados, utilizar CrawNet para realizar búsquedas de un tema en particular utilizando como plantillas de búsqueda los formularios generados.

REFERENCIAS

[1] H. Yeye, X. Dong, G. Venkatesh, R. Sriram & S. Nirav, “Crawling Deep Web Entity Pages”, Proceedings of the Sixth ACM International Conference on Web Search and Data Mining. Rome, Italy, pp. 355-364, 2013

[2] M. Bergman, “White Paper: The Deep Web Surfacing Hidden Value”. BrightPlanet: The Journal of Electronic Publishing, vol. 7, no. 1, 2012.

[3] *M. Álvarez.* “Arquitectura para Crawling dirigidode información contenida en la web oculta”. PhD. Dissertation, *Universidad Coruña*, A. Coruña, España, 2007.

[4] S. Lawrence & C. Giles, “Accessibility of Information on the Web”, *Nature*, vol. 400, no. 1, pp. 107-109, Julio, 1999.

[5] B. Fernández. & S. Pardo, “Selección de recursos de información disponibles en el Web invisible”. *Acimed*. vol. 14, no. 6, 2006.

[6] Z. Wu, L. Jiang, Q. Zheng & J. Liu, “Learning to surface deep web content”, In Proc. 2011, Twenty-Fourth AAAI Conference on Artificial Intelligence. *Georgia, USA, pp. 1967-1968.*

[7] W. Yan, “Query selection in deep web crawling: Help your crawler efficiently retrieve data from the largest data sources in the web year”, Scholar’s Press, 2014.

[8] K. Chang, B. He, & Z. Zhang, “Toward large scale integration: Building a MetaQuerier over databases on the web”, 2005. Proceedings of the Second Biennial Conference on Innovative Data Systems Research, Asilomar, CA, USA, January, pp. 44-55.

[9] B.He, M. Patel, Z. Zhang & K. Chen-Chuan, “Accessing the deep web: A survey”, Commun. ACM, vol. 50 no. 5, pp 94-101. Mayo, 2007.

[10] M. Soulemane, M. Rafiuzzaman & H. Mahmud, “Crawling the hidden web: An approach to dynamic web indexing”, *International Journal of Computer Applications,* vol. 55, no. 1, pp 7-15, Octubre, 2012.

[11] D. Anuradha & A. Babita, “Hidden web extractor: Dynamic way to uncover the deep web”, *International Journal on Computer Science & Engineering,* vol. 4, no. 6, pp. 1137-1145. Junio, 2012.

[12] L. Xian, D. Xin, L. Kenneth, M. Weiyi & S. Divesh, “Truth finding on the deep web: Is the problem solved?”, Proceedings of the *39th international Conference on Very Large Data Bases,* Trento, Italy, pp. 97-108, 2013

[13] S. Liddle, S. Yau & D. Embley, “On the automatic extraction of data from the hidden web”. In Proc. 2001 *International Workshop on Data Semantics in Web Information Systems (DASWIS-2001),* London, UK, pp. 212-226.

[14] C. Sherman & G. Price. 2001, “The invisible web: Uncovering information sources search engines can’t see”. Medford, N.J, CyberAge Books, 2001.

[15] M. Álvarez, J. Raposo, A. Pan, F. Cacheda, F. Bellas, & V. Carneiro. “Crawling the content hidden behind web forms”, Proceedings of the ICCSA, Lecture Notes in Computer Science v 4706, Springer, pp. 322-333, 2007

[16] V. Prieto, M. Álvarez, R. López-García & F. Cacheda, “A scale for crawler effectiveness on the client-side hidden web”, Computer Science and Information Systems, vol. 9 no. 2, pp. 561-583. Junio, 2012.

[17] D. Lewandowski & P. Mayr, “Exploring the academic invisible web”. *Library Hi Tech.,* vol. 24, no. 4, pp. 529-539, Feb. 2007.

[18] M. Wuand & A. Marian, “A framework for corroborating answers from multiple web sources”. *Information Systems,* vol. 36, no. 2, pp. 431-449, Jun. 2011.

[19] X. Dong, B. Saha & D. Srivastava, “Less is more: Selecting sources wisely for integration”. *PVLDB,* vol. 6, no. 2, 2013. Disponible en: http://www.vldb.org/pvldb/vol6/p37-dong.pdf

[20] S. Raghavan, & H. Garcia-Molina, “Crawling the hidden web,” Proceedings of the 27th International Conference on Very Large Data Bases (VLDB 2001), San Francisco, CA, USA, pp. 129-138, 2001

[21] M. Cafarella, E. Chang, A. Fikes, A. Halevy, W. Hsieh, A. Lerner, J. Madhavan & S. Muthukrishnan, “Data management projects at Google”. ACM SIGMOD Record vol. 37, no. 1, pp.34-38, 2008.

[22] Salinas Martínez, Osvaldo. “Modelado semántico de documentos con estructura definida”. Tesis, Cd. Victoria, Tamaulipas, México, Centro de Investigación y de Estudios Avanzados del Instituto Politécnico Nacional, 2012.

[23] F. Martínez-Méndez. Recuperación de información: Modelos, sistemas y evaluación, Murcia, España, Ed. El Kiosko, 2012.