MEI
II, vol. 4
nº 7

Nuevos retos de la tecnología web crawler para la recuperación de información

Manuel Blázquez Ochando
Departamento de Biblioteconomía y Documentación. Universidad Complutense de Madrid

Recibido el
20-12-2013

Aceptado en
06-01-2014

Resumen
El web crawler constituye una parte importante de la cadena documental en la recuperación de información, dado que genera el corpus documental necesario sobre el que aplicar los distintos algoritmos de recuperación. Dada su relevancia, se analiza el papel que desemAlgunas de las conclusiones obtenidas, apuntan a la introducción del reconocimiento del peñan, sus distintos enfoques, aportaciones significativas y estado de la técnica. marcado semántico en la web, al desarrollo de un web crawler más polivalente, capaz de interactuar con la web social y realizar campañas de comunicación.

Palabras clave:
web crawler, recuperación de información, marcado semántico, Apache Nutch, Heritrix, WIRE, SocSciBot, Mbot

New challenges of web crawler technology retrieving information for

Abstract
New challenges of web crawler technology for information retrieval.
The web crawler is an important part of the documentary information retrieval chain, as it generates the necessary documentary corpus on which apply different recovery algorithms. Given their importance, their role, their different approaches, significant contributions and state of the art is discussed. Some of the conclusions point to the introduction of the recognition of semantic markup on the web, the development of a Web crawler most versatile, capable of interacting with the social web and make communication campaigns.

Keywords
web crawler, information retrieval, semantic markup, Apache Nutch, Heritrix, WIRE, SocSciBot, Mbot

1.- Introducción

El perfeccionamiento de los sistemas web crawler, constituye uno de los principales retos en materia de recuperación de información, íntimamente vinculado a los métodos de metadescripción mediante metadatos y marca do semántico. Ello unido a la consecución de herramientas tecnológicas más completas y polivalentes, constituyen algunos de los aspectos que deberán ser resueltos en breve por la comunidad científica y que son pues tos de relieve. En este artículo, se introduce el contexto de la tecnología web crawler en el entorno de la recuperación de información, definiendo los principales enfoques con que éstos son diseñados para finalizar con una reflexión al respecto de los presentes y futuros desafíos a

BLÁZQUEZ OCHANDO, MANUEL. "Nuevos retos de la tecnología web crawler para la recuperación de información". En: Métodos de información (MEI), II Época, Vol. 4, nº, 7, 2013, pp. 115-128.
DOI: http://dx.doi.org/10.5557/IIMEI4-N7-115128

los que la tecnología documental deberá dar respuesta.

2.- Web crawler en el contexto de la recuperación de información

La recuperación de información es la ciencia encargada del tratamiento de la información contenida en una base de conocimiento compuesta por documentos, que son el objeto de consulta por parte de la figura del usuario, dando como resultado aquellos que mejor respondan a sus necesidades informativas y documentales.

La consecución de todos estos objetivos es posible desarrollando procesos que involucran la génesis del corpus documental, su depuración, el análisis estadístico de los términos que constituyen los documentos de la colección, su factor discriminatorio y su peso, el método de almacenamiento de la información, los distintos modelos de recuperación, la representación de la consulta del usuario y de los resultados obtenidos tras su aplicación y la experiencia del usuario que permite retroalimentar el sistema y obtener información valiosa al respecto de la relevancia y precisión con que ha funcionado.

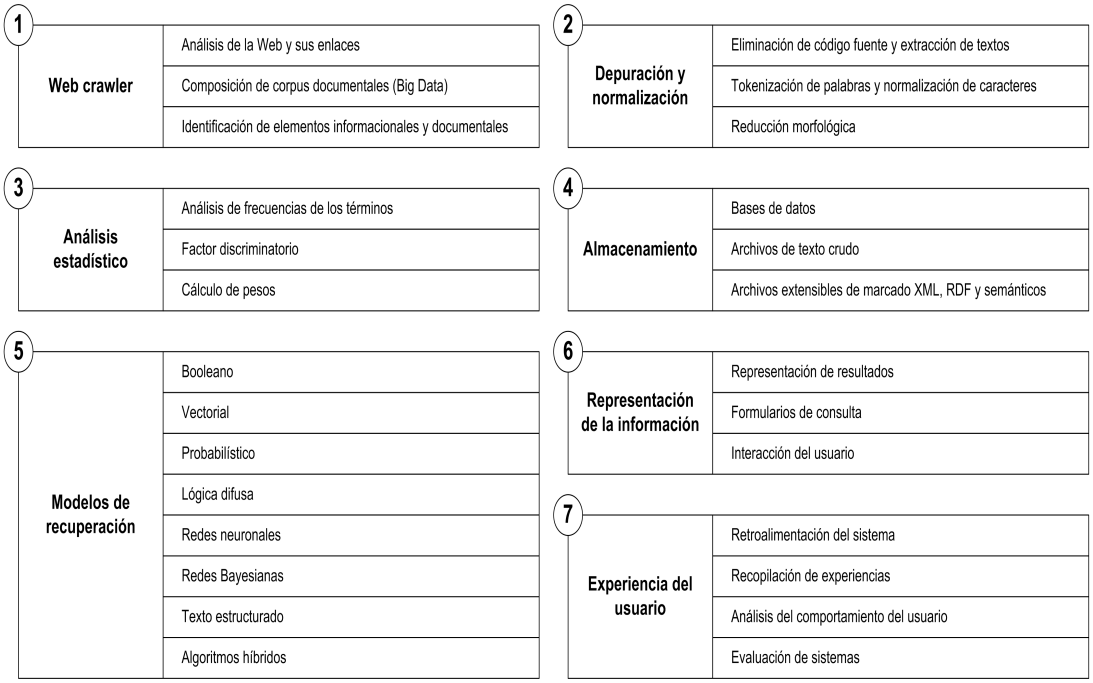

Figura 1. Procesos de la recuperación de información

Sin embargo la base de este proceso lo constituye la colección en sí misma y para ser más precisos, los sistemas que permiten su configuración y recuperación en primera instancia. Estos sistemas son conocidos más comúnmente con el nombre de web crawler y están logrando combinar todas las áreas para las que se diseñaban programas ad hoc. Esto es, unir distintos enfoques y tareas en una sola.

3.- Enfoques de los web crawler

Los web crawler son programas cuya misión principal es el rastreo de páginas web a través de sus enlaces. Este proceso denominado análisis de enlaces es el principal denominador común a todos ellos, puesto que existen especializaciones y enfoques diferentes, que tienen como objetivo la resolución de diversos problemas; 1) indexar la web existente, 2) indexar una parte de la web, 3) extraer recursos de la web, 4) archivar la web para preservarla, 5) realizar estudios webmétricos.

1) Indexar la Web existente. Implica el análisis exhaustivo de la Web a gran es cala, analizando únicamente los vínculos principales entre los sitios y sus páginas, dando prioridad a los enlaces salientes a recursos o dominios externos, frente a los contenidos del propio sitio. El objetivo es obtener todas las relaciones entre sitios y páginas web alojados en servidores y direcciones IP distintas.

2) Indexar parte de la Web. El análisis sectorizado o particionado de la web, im plica la selección a priori de una colección de enlaces denominada semilla, la cual man tiene un común denominador por ejemplo, la temática, el tipo de dominio, recursos, su topografía o cualquier otro factor que sea objeto de estudio. La in dexación par cial de la Web, tiene como objetivo el análisis pormenorizado de los contenidos y elementos de los que están compuestos cada página o sitio web iden tificado en la semilla. El web crawler Mbot (BLÁZQUEZ OCHANDO, M. 2013a) comparte este principio y su nivel de análisis se reduce a la pequeña y media escala.

3) Extraer recursos de la Web. Está vinculado frecuentemente a los análisis de sector de la Web, dada la exhaustividad con la que pueden distinguir la tipología y características de los elementos presentes en los sitios y páginas analizadas. El ob jetivo principal es por tanto la estructuración y ordenación de los contenidos para su posterior tratamiento y aprovechamiento, dando origen al concepto de datami ning Web o minería de datos en la Web.

4) Archivar la Web para preservarla. La continua evolución de la Web como medio de distribución y comunión de contenidos, propicia un ciclo vital de la documentación muy extremo, que conlleva el nacimiento, transformación e incluso muerte de los sitios web y sus informaciones. El cambio, principal constante del movimiento y dinamismo de la Web, presenta el problema de la memoria y por ende, del origen de las fuentes de información y contenidos publicados. Dicho de otra forma, es necesario indexar la Web periódicamente, para permitir el acceso a contenidos pretéritos que bien podrían cambiar en su forma o en su contenido e incluso desaparecer, para no volver a ser recuperados posteriormente. Este es el enfoque del web crawler Heritrix (MARILL, J.L.; BOYKO, A.; ASHENFELDER, M.GRAHAM, L. 2004), empleado para alimentar el portal Internet Archive y recuperar versiones anteriores con el servicio Wayback Machine.

5) Realizar estudios webmétricos. Derivado de los aspectos anteriores, los estudios webmétricos son una consecuencia lógica de analizar la Web cuyo objetivo es su medición cuantitativa. Esto implica el estudio de grafos correspondiente a los enlaces de los sitios web, sus recursos y elementos constituyentes, el análisis de dominios, formatos, metadatos, meta-etiquetas, coenlaces, macroestructura web y ranking de contenidos más vinculados. Por otra parte, también pueden considerarse los estudios del corpus lingüístico de las páginas web analizadas, determinando por ejemplo el análisis de frecuencias, de los términos, factor discriminatorio, simulaciones de reducción morfológica y obtención de muestras de vocabulario.

Aportaciones del web crawler Mbot

Teniendo en cuenta estos enfoques, la tendencia en el desarrollo de programas web crawler es la integración de todas las funciones y en la medida de lo posible, de todos los enfoques. En el caso del programa Mbot, se viene trabajando con la idea de crear un web crawler capaz de realizar análisis a pequeña escala de la Web, aportando el mayor nivel de detalle posible. Algunos ejemplos de este concepto son el trabajo del análisis webmétrico de los medios de comunicación brasileños (BLÁZQUEZ OCHANDO, M. 2012a), la prueba de análisis web de la universidad española (BLÁZQUEZ OCHANDO, M. 2013b) y el análisis de la web medioambiental de la Administración Pública en España (RAMOS SIMÓN, L.F.; ARQUERO AVILES, R.; COBO SERRANO, S.; BLÁZQUEZ OCHANDO, M. 2013)

Este mayor nivel de detalle, tiene por objeto satisfacer las demandas del profesional de la información en sus actividades científicas y productivas. Por este motivo, Mbot diferencia en su esquema de datos, los metadatos, meta-etiquetas, enlaces a páginas web, enlaces a imágenes, enlaces a documentos, enlaces audiovisuales, enlaces sociales, enlaces a ontologías, enlaces a documentos de web semántica RDF, texto indexado, código fuente completo y canales de sindicación. Partiendo de la correcta distinción de estos recursos, se procesan de manera estructurada para su posterior aprovechamiento a modo de Big Data (BARRANCO FRAGOSO, R. 2012) en el contexto de la minería de datos y de los resultados cibermétricos.

Por otra parte, es reseñable la aportación orientada al descubrimiento de canales de sindicación que pueden ser directamente integrados en programas de agregación. Esta tarea, también viene siendo contemplada por la comunidad científica (O'RIORDAN, A.; O'MAHONEY, O., 2011) y resulta de utilidad para alimentar buscadores especializados como Medworm, prestar servicios de información en portales de noticias como Google News o realizar estudios sobre la producción informativa de los medios de comunicación de España y México (BLÁZQUEZ OCHANDO, M. 2012b)

Otra aplicación disponible en Mbot, permite utilizar los correos electrónicos obtenidos durante el análisis, para enviar mensajes instantáneos. El programa proporciona una plataforma para la edición de mensajes en formato HTML y administrar de forma sencilla archivos adjuntos. La disponibilidad de esta herramienta hace posible la elaboración de campañas de marketing y publicidad adaptada a la semilla analizada por el webcrawler. La sectorización de la web, permite afinar en mayor medida el público objetivo, proporcionando una vía de difusión que de otra forma, podría resultar costosa de obtener.

Estado de la técnica

Con la intención de determinar qué aspectos son mejorables en los actuales web crawler, se pueden comparar los cuatro principales rastreadores de código abierto, Apache Nutch, Heritrix, WIRE (CASTILLO, C.; BAEZA YATES, R. 2005) y SocSciBot (THELWALL, M. 2012) con el web crawler experimental Mbot, obteniendo así una idea del estado de la técnica, véase tabla 1. Si bien el programa Apache Nutch consta de mayor escalabilidad (KHARE, R.; CUTTING, D.; SITEKAR, K.; RIFKIN, A. 2005), al igual que Heritrix y WIRE, éstos tienen una contrapartida importante, que es la dificultad de su correcta instalación y configuración para realizar las mismas tareas que podrían llevar a cabo programas más sencillos, como SocSciBot y Mbot. Por tanto un reto para el diseño de futuros web crawler es lograr la máxima sencillez de instalación, configuración y puesta en marcha. Dicho de otra forma, conseguir una mayor usabilidad de los programas y en la medida de lo posible, adaptarlos a las necesidades del documentalista y por extensión del científico e investigador.

Por otra parte una aplicación todo en uno, que incorpore con su instalación todos los elementos necesarios para funcionar de forma completa, constituye otro aspecto deseable. En este apartado, Nutch y WIRE requieren cinco extensiones distintas para poder buscar, introducir un interfaz gráfico, indexar, representar gráficos o recuperar información en archivos XML mediante parser. Sin embargo, tanto Apache Nutch como Heritrix, WIRE y SocSciBot integran un método de almacenamiento de datos de tipo noSQL, que mejora la velocidad de ejecución del web crawler (GANESH, T.V. 2012), en detrimento de la eficiencia estructural del sistema de base de datos MySQL que utiliza Mbot.

Tabla 1. Comparativa de sistemas web crawler con enfoques diferenciados

	Nutch	Heritrix	WIRE	SocSciBot	Mbot
Enfoque	Análisis de la Web a gran escala	Archivo de la Web	Análisis de la Web a gran escala	Análisis de enlaces y corpus lingüístico	Análisis de la Web sectorizada
Tipo de software	Libre	Libre	Libre	Libre	Experimental
Distribución	Libre	Libre	Libre	Libre, no comercial	Limitada
Soporte	Comunidad de desarrolladores	Comunidad de desarrolladores	Center for Web Research (Chile)	Statistical Cybermetrics Research Group (University of Wolverhampton)	Autor
Aplicación todo en uno	-	X	-	-	X
Extensiones relacionadas	Apache Hadoop Apache Gora Apache Tika Apache Solr Apache Lucene	-	SWISH-E JWire WIRE-Nic ConNeCTOR GNU plot	SocSciBot Tools Cyclist	-
Compatible con todos los S.O	Sólo Linux	Linux recomendado	Sólo Linux	Windows recomendado	Todos
Tecnología y dependencias	Java 1.6 Virtual Machine + Lista de dependencias	Java 5.0 JRE + Lista de dependencias	Java 1.6 Virtual Machine + Lista de dependencias	Microsoft .NET Framework	Apache Web Server , PHP5 y MySQL
Configuración de todas las opciones del web crawler con interfaz gráfico	-	-	-	-	X
Dificultad de la instalación	Muy alta	Alta	Alta	Baja	Baja
Dificultad de la configuración	Muy alta	Media	Alta	Media	Baja
Varios procesos de análisis simultáneos	X	X	X	X	X
Extracción de enlaces internos y externos	X	X	X	X	X
Reconocimiento de enlaces dentro de códigos JAVA	-	X	-	-	-
Reconocimiento de enlaces dentro de documentos ofimáticos	-	X	-	-	-

	Nutch	Heritrix	WIRE	SocSciBot	Mbot
Extracción y reconocimiento de enlaces a redes sociales	-	-	-	-	X
Extracción y reconocimiento de documentos ofimáticos	X	X	X	X	X
Extracción y reconocimiento de canales de sindicación	Requiere extensión ROME	-	-	-	X
Extracción y reconocimiento de correos electrónicos	Requiere extensión SOLR	-	-	-	X
Extracción y reconocimiento de imágenes y archivos multimedia	X	X	X	X	X
Extracción de metadatos y meta-etiquetas	Requiere extensión TIKA	-	X	X	X
Extracción de código HTML	X	X	X	X	X
Almacenamiento	NoSQL	NoSQL	NoSQL	NoSQL	MySQL
Indexación en BD	-	-	-	-	X
Esquema de datos	type, content, parsedData, parseText, outlinks, metadata, type, namespace	software version, ip host, host name, crawl operator, user-agent, http requested, http header, policy, timestamp, parsed data	text and html storage, metadata index, url index, link index, harvest index	?	domain, url1, url2, date, updated, nlevel, rankmbot, title, metatag, metadata, linkmap, linksync, linkont, linksem, linkdoc, linksoc, linkimg, linkaud, linkvid, linkmail, sourcecode, indexer
Depuración de textos	Requiere extensión Lucene	-	X	X	X
Normalización de caracteres de los textos	Requiere extensión Lucene	-	X	X	X
Reducción morfológica de las palabras	Requiere extensión Lucene	Requiere implementación de código PERL Snowball	-	X	-
Eliminación de palabras vacías	Requiere extensión SOLR	Requiere modificación de archivos	X	X	X

En lo que identificación, reconocimiento, extracción de enlaces y elementos de la web se refiere, destaca la capacidad de distinción y estructuración de Mbot, que almacena cada categoría de recursos con un esquema de datos muy amplio. Sin embargo, la capacidad de detección de enlaces de Heritrix es ligeramente superior, ya que permite la identificación de enlaces sitos en códigos JAVA, CSS y documentos ofimáticos, permitiendo una mayor penetración, propia de su enfoque objetivo: la preservación de la Web y sus contenidos (MOHR, G.; STACK, M.; RANITOVIC, I.; AVERY, D.; KIMPTON, M. 2004). Finalmente de los cinco web crawler, sólo dos no disponen de herramientas de análisis webmétrico incorporadas. Éstos son Apache Nutch y Heritrix, que requieren de desarrollos aislados para poder extraer información útil relativa al análisis de enlaces. Los web crawler WIRE, SocSciBot y Mbot, contemplan instrumentos e informes cuantitativos que permiten caracterizar la web adecuadamente, tal como se muestra en la tabla 2.

Tabla 2. Informes automáticos generados por los web crawler

Análisis e informes automáticos, generados por el web crawler	WIRE	SocSciBot	Mbot
Análisis de datos general, nivel por nivel de la colección	X	X	X
Análisis de dominios de primer y segundo nivel	X	X	X
Análisis general de formatos de archivo	X	X	X
Análisis general de enlaces internos y externos	X	X	X
Análisis de enlaces internos y externos dominio por dominio	X	X	X
Ranking de páginas web más enlazadas	X	X	X
Ranking de sitios web más enlazados	X	X	X
Análisis de enlaces de la web social			X
Meta-etiquetas por dominio			X
Texto de meta-etiquetas por dominio y página			X
Análisis de frecuencias TF de meta-etiquetas			X
Metadatos por dominio			X
Texto de metadatos por dominio y página			X
Análisis de frecuencias TF de metadatos			X
Términos más frecuentes de la colección	X		
Profundidad de la web nivel por nivel	X		X
Ranking de sitios web con más páginas web	X	X	X
Ranking de sitios web con más contenidos y documentos	X	X	X
Ranking de sitios web con más canales de sindicación			X

Análisis e informes automáticos, generados por el web crawler	WIRE	SocSciBot	Mbot
Exportación de canales de sindicación			X
Análisis de la macroestructura de la web	X		X
Análisis de coenlaces	X	X	X
Exportación de archivo gráfico de tipo DOT			X
Generación de diagramas gráficos de la web analizada	X	X	X
Cálculo de pagerank	X		
Cálculo de HITS	X		
Estadísticas de idiomas	X		
Validación de páginas HTML		X	
Archivos de datos crudos para la exportación	X	X	X

Se observa que todos tienen en común los análisis de enlaces, co-enlaces, dominios, formatos de archivo, ranking de páginas, sitios más enlazados y con más contenidos, generación de gráficos de la web y exportación de datos en archivos crudos. Por otra parte existen características especiales de cada programa, por ejemplo SocSciBot realiza pruebas de validación de las páginas web analizadas, WIRE puede calcular el índice Pagerank y HITS de las páginas web analizadas y Mbot permite el análisis en profundidad de los metadatos y meta-etiquetas de la web. Si se toman en consideración estos datos, un reto actual y futuro es el desarrollo de los análisis y estadísticas webmétricas, tanto para lograr equiparar las prestaciones entre programas web crawler, como para ampliar su objeto de estudio.

4.- Nuevos retos en el desarrollo de sistemas web crawler

Habida cuenta de las características y factores que operan en los programas web crawler, se destacan los siguientes retos; 1) simplificar el uso de los web crawler, 2) convertir el web crawler en una herramienta polivalente, 3) desarrollar la capacidad de reconocimiento y aprovechamiento semántico, 4) ampliar la capacidad de difusión.

1) Simplificar el uso de los web crawler. Los programas web crawler no suelen ser sencillos de instalar y menos de configurar. Ello es debido en parte a la plata forma y lenguaje de programación con el que fueron diseñados, dejando fuera de la ecuación al usuario menos experimentado. Es necesario recordar que el desarrollo y manejo de tales programas, no son únicamente objetos de investigación, sino herramientas de las que cualquier investigador podría y debería

servirse. En este sentido podría hablarse de la necesidad de "democratizar" su empleo, simplificando sus mecanismos y sirviendo a un propósito cada vez más polivalente.

2) Convertir el web crawler en una herramienta polivalente. Unido a la simplificación del uso de los web crawler, se encuentra el reto más complicado; conseguir que el web crawler sea capaz de realizar múltiples tareas y procesos para el que se necesitarían diversas aplicaciones:

a. Conseguir que un web crawler se convierta en una plataforma de marketing, capaz de identificar distintos sectores de la web y con ello distintos públicos objetivos, a los cuales transmitir un determinado mensaje.

b. Realizar difusión selectiva de la información a través del análisis de los textos publicados en las web analizadas.

c. Vincular el rastreo de la web con la extracción de textos y la publicación automática de contenidos según categorías temáticas.

d. Realizar estudios webmétricos sectoriales cada vez más avanzados y completos.

e. Depurar procesos automáticos de selección de la información para la toma de decisiones.

f. Incorporar interfaz de búsqueda en tiempo real sobre los contenidos indexados.

g. Capacidad de exportación de datos estructurados.

h. Alimentación de agregadores mediante el descubrimiento de canales de sindicación.

3) Desarrollar la capacidad de reconocimiento y aprovechamiento semántico. El éxito de la web semántica, bien depende de la capacidad con que los web crawler sean capaces de reconocerla y aprovecharla. Esto significa que no todos los web crawler están preparados para detectar su uso y menos indexar adecuadamente sus contenidos para realizar las inferencias semánticas oportunas. En consecuencia pueden y deben plantearse diversas preguntas al respecto ¿Cuántos métodos existen para construir una web semántica? ¿Qué método es más aprovechable? ¿Cómo detectará y recuperará el web crawler los contenidos semánticos? ¿Cómo procesará y almacenará el contenido semántico? ¿Cómo indexará el contenido semántico para su recuperación? ¿Será necesario introducir un motor de inferencia semántica dentro del propio web crawler? ¿Cómo se visualizará la información semántica y cómo ayudará al usuario a tomar decisiones sobre lo que desea buscar?

a. ¿Cuántos métodos existen para construir una web semántica?

Un web crawler debería tener en consideración los principales métodos de construcción semántica con la finalidad de identificar y recuperar todos sus contenidos y respetar sus relaciones. Según las especificaciones oficiales de RDF, el método por defecto es la codificación de un archivo en dicho formato, vinculado a la página web. También existe el enfoque de RDF-A con el que es posible integrar la web semántica en una página web XHTML. Otra opción posible es el empleo de microformatos que actúan directamente en el código HTML.

b. ¿Qué método es más aprovechable?

El método más adecuado sería aquel que permita establecer relaciones semánticas para los distintos fragmentos de texto de una página web, de forma tal que simple fique al máximo los procesos de reconocimiento e identificación del web crawler y de edición del contenido por parte del usuario. Si se pretende divulgar el uso de la web semántica, su método de codificación debería aspirar a ser tan sencillo como el del propio lenguaje HTML, de forma que se generalizara su conocimiento y la construcción de páginas web semánticas fuera sólo una rutina. El método que más se acerca a esta premisa es RDF-A y el empleo de microformatos.

c. ¿Cómo detectará y recuperará el web crawler los contenidos semánticos?

Los web crawler emplean métodos de análisis o parser que actúan sobre el código fuente de las páginas web. El reconocimiento semántico, implica la adición de nuevos métodos de detección y recolección de atributos para RDF-A y microformatos, así como de modelado de estructuras RDF. Parte de la solución podría encontrarse en DOM también conocido como modelo de objetos de documento, con el que es posible organizar jerárquicamente todas las etiquetas tanto de un documento HTML como XML y guardarlas en un Array u objeto manipulable en alguno de los principales lenguajes de programación. Teniendo acceso completo a la estructura de etiquetas y elementos del documento, sea semántico o no, se pueden recuperar la etiqueta o contenidos deseados mediante selección XPath, obteniendo de esta forma los triples del modelo de datos semánticos empleados.

d. ¿Cómo procesará y almacenará el contenido semántico?

Si bien el reconocimiento es indispensable, el procesamiento de cara al almacenamiento de la información resulta fundamental para su aprovechamiento. En un modelo de base de datos, será necesario relacionar la página web analizada con los triples obtenidos y no debería ser descartable el uso de una nueva tabla de la base de datos dedicada a la construcción de un BSD o big semantic document, que aglutinara todos los triples de todas las páginas web analizadas con la finalidad de facilitar su inferencia en motores de búsqueda.

e*. ¿Cómo indexará el contenido semántico para su recuperación?*

El empleo de tablas para la construcción de un BSD puede facilitar la indexación de los objetos de los triples registrados, aunando lo mejor de la recuperación a texto completo de bases de datos como MySQL con la capacidad de relación de la web semántica.

f. ¿Será necesario introducir un motor de inferencia semántica dentro del propio web crawler?

La mayoría de los sistemas web crawler incorporan algún interfaz de recuperación y consulta que permite en mayor o menor medida el testado de los contenidos indexados. Por tanto es lógico pensar, que los motores de inferencia semántica se desarrollen bajo el auspicio de los web crawler, por ser las herramientas que en primera instancia están llamadas a identificar y reconocer sus contenidos.

g. ¿Cómo se visualizará la información semántica y cómo ayudará al usuario a tomar decisiones sobre lo que desea buscar?

Es muy probable que la apariencia de los buscadores semánticos sea semejante a la de los actuales, excepto por la aparición de sugerencias y refinamientos predeterminados por la web semántica razonada. De esta forma, no sólo se proporcionará la información que satisface el modelo de recuperación indicado, sino que se indicarán alternativas de recuperación y navegación entre los resultados relacionados con el objeto consultado, en base a la estructura de triples previamente almacenada en la base de datos.

4) **Ampliar la capacidad de difusión.** Aunque los web crawler no son herramientas diseñadas para realizar una difusión de información, sí son fácilmente adaptables a tales propósitos. La capacidad para extraer las direcciones de correo electrónico de un determinado análisis y emplearlas para transmitir mensajes desde la misma aplicación, puede ahorrar tiempo y facilitar el desarrollo de planes de marketing, promoción y difusión. Pero además de estos recursos, la diferenciación de los enlaces de redes sociales, permitiría la implementación de comunicación y difusión masiva de mensajes en todas ellas. Dicho de otra forma, no resultaría complicado desarrollar un gestor de mensajes para redes sociales que automatizara su publicación de acuerdo a sus patrones y características.

5.- Conclusiones

La complejidad de los procesos de instalación, configuración y la interoperabilidad de los sistemas, así como de dependencias y extensiones, dificulta el uso de los web crawler. Resulta recomendable diseñar sistemas mejor adaptados y depurados para realizar las complejas tareas de rastreo de la web.

El empleo de sistemas de almacenamiento de datos NoSQL permite una velocidad de procesamiento y capacidad mayor que el empleo de bases de datos MySQL. Pero al renunciar a la estructuración de las bases de datos relacionales, también se renuncia a un método sencillo de organización de la información. En consecuencia es recomendable que los nuevos desarrollos permitan la elección del método de almacenamiento que mejor se adapte a las necesidades del usuario.

El futuro de los sistemas de web crawler pasa por convertirse en herramientas multipropósito, capaces de adaptarse mejor a las necesidades y realidades de cada

usuario. Esto significa sencillez de uso y adecuación tanto a las necesidades del mismo, como a las posibilidades que brinda la recuperación de información y contenidos de la Web. Ello quiere decir, la incorporación de funciones automatizadas para la elaboración de estudios webmétricos, plataformas para la gestión de campañas de publicidad y marketing social y correo electrónico, sistemas de datamining para el aprovechamiento de fuentes de información filtradas, herramientas de vigilancia informacional para monitorizar los datos recuperados y de ayuda para la toma de decisiones.

Los sistemas web crawler están orientados al análisis de los contenidos hipertextuales y documentales pero no a la recuperación de contenidos semánticos. Esto es la recuperación mediante inferencia de sus relaciones y no la mera detección de los archivos semánticos codificados en RDF.

Para lograr una perfecta adecuación entre la web semántica, el marcado semántico y los sistemas de recuperación de información basados en web crawler, se hace necesario un nuevo enfoque para codificar y establecer relaciones semánticas de una forma más sencilla e intuitiva. En este sentido es necesario pensar en una nueva web semántica que sea tan sencilla en su edición, que cualquier editor de la web con conocimientos básicos de HTML pueda fácilmente codificarla.

6.-Bibliografía

BARRANCO FRAGOSO, R. 2012. ¿Qué es Big Data? En: Information Management, Biblioteca técnica. IBM. Disponible en: http://www.ibm.com/developerworks/ssa/local/im/que-es-big-data/index.html?cmp=BS&ct=SocialMedia&cr=twitter

BLÁZQUEZ OCHANDO, M. 2012a. Análisis webmétrico de los medios de comunicación brasileños: prensa, radio y televisión. En: I Seminario Hispano Brasileño de Biblioteconomía y Documentación (Madrid, 28-30 noviembre). Disponible en: http://eprints.rclis.org/19033/

BLÁZQUEZ OCHANDO, M. 2012b. Desarrollo de un sistema de clasificación automática de contenidos en medios de comunicación españoles y mexicanos. En: 9º Seminario Hispano-Mexicano de Biblioteconomía y Documentación (México, 7-9 mayo). Disponible en: http://eprints.rclis.org/19031/

BLÁZQUEZ OCHANDO, M. 2013a. Mbot: Webcrawler multipropósito. Disponible en: http://mblazquez.es/mbot/

BLÁZQUEZ OCHANDO, M. 2013b. Desarrollo tecnológico y documental del webcrawler Mbot: prueba de análisis web de la universidad española. En: XIII Jornadas Españolas de Documentación, FESABID (Toledo, 21-24 mayo). Disponible en: http://eprints.rclis.org/20404/

CASTILLO, C.; BAEZA YATES, R. 2005. Wire: an open-source web Information Retrieval environment. En: Workshop on Open Source Web Information Retrieval (OSWIR). pp. 27-30. Disponible en: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.81.6859&rep=rep1&type=pdf#page=27

GANESH, T.V. 2012. When NoSQL makes better sense tan MySQL. En: The Tech Trek. IBM. Disponible en: https://www.ibm.com/developerworks/community/blogs/theTechTrek/entry/when_nosql_makes_better_sense_than_mysql8?lang=en

GIBBNEY, L.J. 2012. About Plugins. En: Nutch Wiki. Disponible en: http://wiki.apache.org/nutch/AboutPlugins

GIBBNEY, L.J. 2013. Plugin Central. En: Nutch Wiki. Disponible en: http://wiki.apache.org/nutch/PluginCentral

KHARE, R.; CUTTING, D.; SITEKAR, K.; RIFKIN, A. 2005. Nutch: A Flexible and Scalable Open-Source Web Search Engine. Oregon State University. p 32. Disponible en: http://www.master.netseven.it/files/262-Nutch.pdf

LASKOWSKI, S. 2005. Web Metrics Testbed. National Institute of Standards and Technology. Disponible en: http://zing.ncsl.nist.gov/WebTools/

MARILL, J.L.; BOYKO, A.; ASHENFELDER, M. GRAHAM, L. 2004. Tools and Techniques for Harvesting the World Wide Web. En: Proceedings of the 2004 Joint ACM/IEEE Conference on Digital Libraries (JCDL'04). Disponible en: http://www.computer.org/csdl/proceedings/jcdl/2004/2493/00/24930403.pdf

MOHR, G.; STACK, M.; RANITOVIC, I.; AVERY, D.; KIMPTON, M. 2004. An Introduction to Heritrix: An open source archival quality web crawler. En: 4th International Web Archiving Workshop. Disponible en: http://iwaw.europarchive.org/04/Mohr.pdf

O'RIORDAN, A.; O'MAHONEY,O. 2011. Engineering an Open Web Syndication Interchange with Discovery and Recommender Capabilities. Journal of Digital Information, 12 (1). Disponible en: http://journals.tdl.org/jodi/index.php/jodi/article/view/962/1744

RAMOS SIMÓN, L.F.; ARQUERO AVILÉS, R.; COBO SERRANO, S.; BLÁZQUEZ OCHANDO, M. 2013 La información medioambiental en España: recursos y acceso a la información pública (1ª Parte). En: Revista Interamericana de Bibliotecología, 36 (3). pp. 221-234. Disponible en: http://aprendeenlinea.udea.edu.co/revistas/index.php/RIB/article/view/17981/15468

THELWALL, M. 2012. SocSciBot 4. Statistical Cybermetrics Research Group. Disponible en: http://socscibot.wlv.ac.uk/index.html