Método para la extracción masiva de canales de sindicación

A method for the massive extraction of syndication channels

Manuel BLÁZQUEZ OCHANDO

Departamento de Biblioteconomía y Documentación, Universidad Complutense de Madrid, España, manu-blaz@ucm.es

Resumen

Uno de los problemas para la investigación de la producción informativa de canales de sindicación es contar con la cantidad de fuentes suficientes y del mismo dominio, temática o área de conocimiento, para componer una muestra. Esto se debe a la dispersión de las fuentes de información en la Web y por otra parte a la dificultad del investigador para conocer todos los recursos disponibles. A estos problemas se suma la dificultad de extraer y localizar los enlaces de los canales de sindicación en cada recurso o sitio web pertinente que se descubre. En este artículo se aborda el método para extraer los canales de sindicación mediante la composición de semillas para el análisis, con programas *web crawler*, identificando la configuración y posterior preparación.

Palabras clave: Minería de datos. Extracción de datos. Web crawler. Sindicación de contenidos. RSS. Canales de sindicación.

Abstract

One of the problems for investigating the informative production of syndication channels is counting on the sufficient number of sources from the same domain, subject or area of knowledge, to compile a sample. This is a consequence of the dispersion of information sources on the Web; the researcher's difficulty in knowing all the available resources; and the difficulty in extracting and locating the links of syndication channels in every relevant web site or Internet resource that is discovered. This article discusses the method to extract and compile syndication channels through the composition of seeds using a web crawler, and the configuration and subsequent processing of the obtained links.

Keywords: Data mining. Scraping. Web crawler. Content syndication. RSS. Feeds.

1. Introducción

La redifusión de información, sinónimo de sindicación de contenidos, corresponde a una tecnología clave para soportar diversas actividades de la información y documentación. De hecho, la sindicación se define como la transmisión de activos informativos y documentales, para su reutilización e integración en terceros recursos, a través de un archivo editado en lenguaje de marcado extensible XML, contenedor de la información, conforme a un formato de estructuración de datos RSS o Atom. A este recurso se le denomina canal de sindicación (Hammersley, 2003). A su vez, los canales de sindicación, normalizados de acuerdo a un formato o sintaxis determinada, pueden ser suscritos por los usuarios o destinatarios, mediante programas de agregación, especializados en la lectura y gestión de dichos canales. Ello ha permitido que las Ciencias de la Documentación aprovechen la tecnología para la redifusión de registros documentales y autoridades, el intercambio de contenidos entre bibliotecas digitales (Eito-Brun, 2015, p. 218), la alimentación de grandes proyectos europeos como Europeana (Houssos et al., 2011, p.2), la difusión selectiva de la información y su aplicación como sistema de vigilancia informacional y tecnológica (Peis et al., 2008, p. 522) o el desarrollo de estrategias de redifusión social y curación de contenidos (Nieto, 2015, p. 5). Desde un punto de vista informativo, la sindicación tiene un papel fundamental para el periodismo digital, como herramienta de difusión de los medios, como tecnología democratizadora del derecho de acceso a la información. Ello se debe a la libertad de elección del lector sobre las fuentes de información públicas, permitiéndole discriminar sus contenidos, comentarlos y sugerirlos, conformando así su propio menú de informaciones (McCownet al., 2009, p. 1-2).

Como puede deducirse, la sindicación de contenidos está presente en los procesos de comunicación más importantes y ello ha propiciado una explosión demográfica en la creación de canales de sindicación, que está determinada por el uso de los sistemas de gestión de contenidos digitales. Si se asume que existen 348 millones de instalaciones de programas CMS registradas (BuiltWith, 2016), con sus correspondientes canales de sindicación, sin incluir sitios gubernamentales, instituciones académicas, bibliotecas y medios digitales. Se concluye que el principal problema es el descubrimiento de los canales de sindicación pertinentes. A ello se añade la dificultad de

recuperar masivamente la información, crear *big data*, clasificar y organizar los resultados, a fin de evaluar e integrar los canales de mayor valor en los flujos de comunicación. En esta investigación se plantean algunos interrogantes en relación a este problema: ¿Cómo detectar los canales de sindicación de un determinado dominio o área de conocimiento? ¿Qué métodos y herramientas existen para recuperar de forma masiva los canales de sindicación? ¿Cómo se discriminan los formatos de sindicación defectuosos?

El método que se propone combina programas web crawler y estrategias de búsqueda para dirigir el objetivo de recopilación. Si bien el uso de estos métodos de forma aislada es conocido, no lo es tanto a la hora de definir un método cualitativo para la extracción masiva de canales de sindicación, esto es, los enlaces a los recursos en formato RSS.

Las últimas investigaciones en la materia se refieren a técnicas de rastreo web para mejorar la calidad del contenido de los canales de sindicación mediante la implementación de contenidos extras relacionados (Hurtado, 2015, p. 2-3), buscar contenidos especializados en una fuente de información específica (Lubbers, 2015, p. 21-26), desarrollar agregadores de contenidos (Baporikar et al., 2015, p. 159-160), e incluso el uso de programas web crawler especializados con capacidades de consulta adaptativa (Lee et al., 2008, p. 220-223) destinados a la recopilación de noticias producidas en estos medios. En la misma línea, también es destacable la experiencia de buscador especializado en contenidos producidos por canales RSS, elaborado por Viseur (2012).

Sin embargo, las referencias citadas no abordan una técnica con la que puedan ser descubiertos nuevos canales de sindicación, desconocidos para el investigador en una materia o temática dada. Esto es, los enlaces de archivos RSS, con los que poder crear una colección lo suficientemente amplia como para ser considerada del dominio del Big Data.

La extracción masiva de canales de sindicación en un dominio concreto, permitiría estudios más exhaustivos y precisos, en relación a la producción informativa de los medios de comunicación de un país, de las revistas científicas en un área de conocimiento, de la universidad, la administración pública e incluso las redes sociales. De hecho, otros investigadores ya advirtieron la posibilidad de usar la información que los canales de sindicación proporcionan, para identificar debates públicos en los medios, determinar su relevancia, así como el grado de preocupación de la opinión pública (Thelwall et al., 2006, p. 1-3).

También cabe mencionar el papel de la sindicación de contenidos frente a herramientas de alerta de noticias como Google Alerts, cuyo objetivo es proporcionar una vigilancia informacional basada en las palabras clave. El resultado es el suministro de alertas de noticias procedentes de medios y fuentes específicas tales como vídeos, finanzas, libros y blogs. Sin embargo, no es posible conocer a priori, qué fuentes de información está utilizando y sobre todo la calidad de las mismas. Este hecho ya fue descubierto en investigaciones pretéritas (Prieto, Lloret, y Palomar, 2012) en las que se concluye que la efectividad de este tipo de herramientas dependía de los términos de consulta empleados y del dominio general de las alertas, que no siempre se adecua al área de conocimiento abordada. Este tipo de experiencias apoya la tesis de controlar y descubrir nuevas fuentes de información especializadas en la materia, frente al dominio global de la web y el desarrollo de estrategias de consulta basadas en lenguajes documentales con un cierto grado de normalización. Es en este apartado donde las técnicas de extracción y descubrimiento de canales de sindicación pueden ayudar al investigador a desarrollar sistemas de vigilancia informacional más certeros. Por todo ello, los canales de sindicación pueden tener aplicaciones clave en el seguimiento de información especializada, análisis de correlaciones entre contenidos, minería de datos, generación de big data, cálculo de impacto, periodismo de datos o el desarrollo de sistemas expertos para la asistencia en la toma de decisiones.

2. Metodología

El descubrimiento y recopilación masiva de nuevos canales de sindicación puede resultar una tarea complicada, si se tiene en consideración la dificultad de localizar fuentes de información fiables que proporcionen enlaces permanentes a los archivos de datos XML. Si se añade la necesidad de recuperar de forma masiva canales de sindicación para crear nuevas bases de conocimiento o colecciones de investigación, el problema se hace más evidente y obliga a desarrollar una metodología específica. El método propuesto, consta de los siguientes pasos: 1) delimitación del área de conocimiento y desarrollo de un vocabulario representativo, 2) diseño de estrategias de consulta para la recuperación de sitios web pertinentes, 3) creación de una semilla de enlaces para su análisis con herramientas web crawler, 4) análisis de enlaces con programas web crawler, y 5) preparación previa de los canales de sindicación para su procesamiento en agregadores de contenidos.

2.1. Área de conocimiento y vocabulario

La delimitación del área de conocimiento es fundamental para centrar el objetivo de rastreo, que se llevará a cabo posteriormente. Ello implica conformar un corpus textual, representativo del dominio temático en el que se pretenden descubrir nuevos canales de sindicación. El núcleo léxico de las consultas, pasa por la recopilación de descriptores, obtenidos a partir de la literatura del área, la identificación de los temas principales en las revistas científicas, el análisis estadístico de frecuencias de los términos de los textos especializados, así como la revisión de los términos que componen los títulos más citados (Huang et al., 2015, p. 3). La importancia del trabajo de recuperación léxica, reside en la experiencia que el investigador adquiere para componer consultas con terminologías equivalentes o paralelas en la consecución de un objetivo de búsqueda concreto, demostrándose que ello supone una clara ventaja en la recuperación de información (Thatcher, 2008, p. 1309-1311). La organización del léxico no es rígida y basta con definir la categoría de cabecera, las subcategorías principales y sus términos específicos, recogidos a modo de bolsa de elementos, como se puede comprobar en la tabla I.

Esquema de organización léxica
Término cabecera Categoría principal n1 (Término específico n1.1, Término específico n1.2, Término específico n1.3…| Término relacionado n1.1, Término relacionado n1.2, Término relacionado n1.3…)

Ejemplo de organización léxica
Comunicación Secciones de contenidos (Política, Parlamento, Política internacional, Cooperación política, Conflictos, Legislación, Finanzas, Función pública, Derecho civil, Derecho penal, Justicia, Derecho internacional, Libertades, Política económica, Análisis económico, Política comercial, Consumo, Instituciones financieras…)

Definición de calificadores

Calificador	*Descripción*	*Ejemplo*
Geográfico, Idioma	Localización o ubicación del recurso, país o región	Portugal, Alentejo, Algarve, Lisboa, portugués, Pt; Francia, Normandía, Bretaña, Aquitania, Alsacia, francés, Fr
Cronológico	Situación temporal en un año o intervalo específico	2010..2016
Tipo de recurso	Identificación de los tipos de recursos o fuentes de información que las que se desean obtener canales de sindicación	Scientific journal, Paper, E-print archive, Digital repositories, Digital newspaper, Magazines, Mass media, Media, Social network
Editor	Sistema de gestión y edición de contenidos que utiliza el recurso o fuente de información.	Blog, Blogger, Wordpress, Joomla, Drupal, Revist
Extensión y formato	Formato del tipo de archivo XML que se desea recuperar	RSS, Atom, MARC-XML, OAI-MARC, OPML, RDF, OWL
Texto denotativo de enlace	Cadenas de texto frecuentemente presentes en las direcciones URL permanentes de los canales de sindicación	feed, xml, rss, syndication, channel

Tabla I. Esquema de organización léxica y calificadores

Adicionalmente conviene identificar los calificadores, que podrán ser empleados en el diseño de las estrategias de consulta. Los calificadores geográficos permitirán recuperar sitios web de una determinada localización y tendrán implicaciones en el idioma de los recursos. El factor cronológico puede ayudar a diferenciar recursos o fuentes de información obsoletas, o que no reciben actualización desde una determinada fecha, evitando recopilar canales de sindicación muertos. El tipo de recurso es un calificador que ayuda a afinar los resultados en torno a un tipo de publicación web, como por ejemplo revistas científicas, repositorios digitales, medios de comunicación social, etc. La extensión o formato, también ayuda a localizar aquellos sitios web que posean archivos de las características indicadas. El calificador "texto denotativo de enlace", se refiere a ciertas cadenas de caracteres que frecuentemente se encuentran presentes en los enlaces permanentes de algunos sitios web. Ello denota unívocamente la presencia de canales de sindicación. Por ejemplo, las palabras feed y rss se usan en consultas como (inanchor:"feed" OR inanchor:"rss") que recuperarían todos los sitios web, que contengan dichas cadenas de texto.

2.2. Diseño de estrategias de consulta

Una vez preparado el léxico y los calificadores, se define la estrategia de consulta. El objetivo es diseñar consultas optimizadas para buscadores, de los que se extraerán las páginas de resultados más relevantes. De los resultados se obtienen listas de enlaces, con las que conformar una relación más amplia, denominada semilla o matriz. La semilla representará el dominio temático de la web, de la que se desean extraer los canales de sindicación. Por ello las estrategias de consulta (Blázquez, M. 2013a, p. 71-76), requieren operadores avanzados, que aseguren el filtrado efectivo de la web, para aumentar el grado de precisión. En la tabla II se muestra una selección de operadores y ejemplos de restricción de dominio, búsqueda en títulos, corpus textual y enlaces. También se exponen estrategias de consulta para recuperar enlaces de revistas científicas de acceso abierto, en editoriales científicas, medios de comunicación, instituciones científicas y repositorios especializados.

Operador	*Función*	*Ejemplos*
site:	Buscar entre los contenidos del sitio o dominio especificado.	*Buscar revistas científicas de Open Access especializadas en bibliotecas* [site:doaj.org "library"] *Buscar contenidos de Elsevier especializados en medicina* [site:elsevier.com "clinical medicine"]
intitle:	Buscar en los títulos de las páginas web indexadas.	*Buscar secciones de noticias especializadas en economía española* [intitle:noticias intext:españa intext:economía] *Buscar instituciones científicas de Estados Unidos* [intitle:"national institute" intext:"science" intext:"united states"]
intext:	Buscar en el corpus textual de las páginas indexadas.	*Buscar directorios de medios de comunicación especializados en prensa digital en el dominio de Reino Unido* [intext:"directory" intext:"journalism" intext:"newspapers" inurl:uk]
inurl:	Buscar cadena de texto en las direcciones URL indexadas.	*Buscar instituciones científicas de Estados Unidos de tipo gubernamental* [intitle:"national institute" intext:"science" intext:"united states" inurl:gov]
inanchor:	Buscar cadena de texto en los enlaces hipertextuales disponibles en el corpus textual de las páginas web.	*Buscar sitios web que contengan enlaces a repositorios especializados en Biblioteconomía y Documentación* [inanchor:"repository" intext:"library and information science"]

Tabla II. Principales operadores y ejemplos de estrategias de consulta

El diseño de las estrategias de consulta puede expresarse mediante combinaciones de términos, propios del léxico mencionado anteriormente. La consulta puede tener dos posibles orientaciones. La primera consiste en satisfacer una necesidad de información concreta y previamente razonada por el investigador. La segunda es capaz de resolver una necesidad concreta, para aplicar la matemática de combinaciones entre los términos del léxico, con la finalidad de obtener recursos y resultados de la Web que son parcial o completamente desconocidos para el investigador. Esto significa que las estrategias de consulta pueden combinar las necesidades de información puntuales del investigador y añadir la componente matemática de la minería de datos a las consultas.

En este sentido, se propone una fórmula que ayude a configurar una estrategia de consulta combinatoria, basada en la intersección del término cabecera, las categorías principales, los términos específicos y relacionados, así como los calificadores o complementos que ayuden a precisar y filtrar aún más los resultados.

$$Q_n = TC_{n\{0,1\}} \cap CP_{n\{0,2\}} \cap (\text{TE}\leftarrow\rightarrow\text{TR})_{n\{1,4\}} \cap C_{\neq n\{0,3\}}$$

Tabla III. Propuesta de estrategia de consulta

En la tabla III, se identifica que “n” consultas serán el resultado de combinar de forma opcional el término cabecera mediante la intersección opcional de un máximo de dos categorías principales del mismo rango, añadiendo la intersección de uno a cuatro términos específicos o relacionados del mismo rango, y finalmente introduciendo hasta tres calificadores opcionales para el filtrado de las consultas de diferentes rangos

2.3. Creación de una semilla de enlaces

La estrategia de consulta en buscadores proporciona resultados de utilidad para configurar la semilla de enlaces. El problema se encuentra en la extracción de los mismos. Para resolverlo pueden realizarse copias sistemáticas de cada pá-

gina de resultados utilizando algunos complementos de los principales navegadores web (1), ya que los buscadores no posibilitan la descarga libre de sus contenidos, salvo excepciones (2). De hecho, la API del buscador Google para la recuperación de sus páginas de resultados, se encuentra fuera de servicio desde el año 2010 (Google, 2015) y la alternativa ofrecida sólo plantea una solución para crear un buscador de sitio web, que no responde a las necesidades que se están planteando. Otra posibilidad más efectiva, pero menos frecuente, es el empleo de servicios de descarga de enlaces en buscadores. Un ejemplo es la aplicación experimental Google2down (3), diseñada para la recuperación automática de enlaces de múltiples páginas de resultados, en los buscadores Google y Google Scholar.

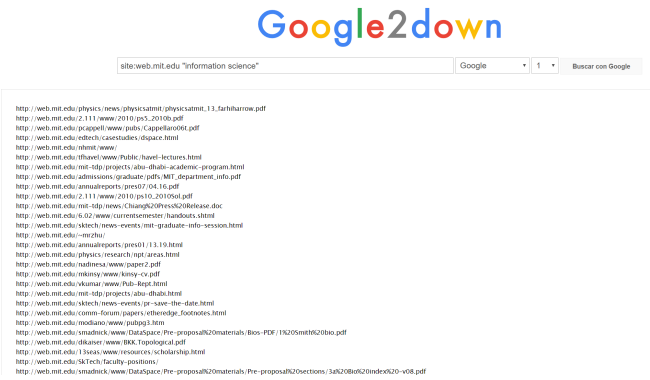

Figura 1. Servicio de descarga Google2down

2.4. Análisis de enlaces con web crawler

Una vez se dispone de la lista completa de enlaces, se procede a la configuración y ejecución de un programa web crawler que los analizará en profundidad a fin de extraer los canales de sindicación. Cualquier web crawler dispone de opciones de configuración que permiten programarlo con distintos niveles de profundidad de análisis y restricción de acceso a dominios y páginas web especificadas. Se aconseja que el número máximo de niveles de profundidad del análisis no exceda el valor 4, ya que está demostrado que más del 70% de los contenidos de la web son recuperables en los tres primeros niveles de análisis (Baeza-Yates, Castillo y Graells, 2008, p. 13). Por ejemplo, un sitio web expresado en la semilla con la designación “p1”, puede contener enlaces a otras páginas web del mismo dominio, designadas como “p1.1, p1.2, p1.n” que a su vez pueden contener enlaces a otras páginas designadas como “p1.1.1, p1.1.2, p1.1.n”. Cada salto de una página web a otra, dentro de un mismo dominio en nivel jerárquico descendente, se denomina nivel de enlazamiento o nivel de profundidad del análisis. La opción “restricción de sitio” debe ser activada, para indicar al web crawler que no tiene que rastrear los enlaces salientes a páginas externas. De esta forma se evita que analice otros sitios web que no siempre tienen que ver con el objetivo planteado en la investigación. Para concluir la configuración, se debe activar la función “extracción de canales de sindicación” o bien definir los patrones de reconocimiento mediante expresiones regulares. Por ejemplo la siguiente expresión regular (<?xml version="…" encoding=".*"?> | <rssversion="…" | <channel> | <ítem>) presenta fragmentos de código que identifican unívocamente cualquier archivo XML y canal de sindicación en formato RSS.

En cuanto al funcionamiento del web crawler, el proceso ya fue descrito en anteriores investigaciones (Blázquez, 2013b) de las que se pueden sintetizar los siguientes pasos: 1) Lectura de los enlaces disponibles en la semilla, 2) Selección del primer enlace, 3) Comprobación de errores de codificación del enlace, 4) Verificación por socket del enlace, mediante envío de cabeceras HTTP Head, vía puerto 80, 5) Verificación de respuesta del servidor, 6) Descarga del código fuente del sitio web enlazado, mediante técnica cURL, 7) Análisis del código fuente y extracción de todos los enlaces de la página, usando el método XPath de selección y expresiones regulares, tal como se muestra en la tabla IV, 8) almacenamiento de los enlaces recuperados en base de datos, 9) replicación del método con toda la semilla de enlaces, 10) cribado de enlaces almacenados, véase apartado 2.5.

Método	*Ejemplo*
XPath	*Extracción de canales RSS* $canalesrss = $xpath -> query("/html/head//link[@type='application/rss+xml']"); *Extracción de enlaces* $enlaces = $xpath -> query("/html/body//a");
REGEXP (Expresiones Regulares)	*Extracción de enlaces* [-a-zA-Z0-9:_\+.~#?&//=]{2,256}\.[^@\][a-z]{2,12}\b(\/[-a-zA-Z0-9:%_\+.~#?&//=]*)?

Tabla IV. Método de extracción de enlaces

2.5. Preparación de canales de sindicación

El resultado del análisis con web crawler suele ser un archivo con la lista de canales de sindicación procedentes del conjunto de sitios web analizados en la semilla. Es muy frecuente recuperar más canales de sindicación de los que realmente corresponden. Ello se debe a que no todos los archivos XML están bien construidos, o bien corresponden a formatos de sindicación en los que

el investigador no está interesado, o bien presentan defectos sintácticos que hacen imposible su lectura (Tabla V). Por estos motivos es fundamental realizar un proceso de cribado.

Principales errores	*Descripción*
(^<\!doctype|^<html>)	Confusión de códigos HTML con códigos XML propios del formato de sindicación
(^\s|^\t|^\R|^\n)	La primera línea XML está precedida de un espacio en blanco que invalida la correcta formación del documento
(rssversion)	El nombre de la etiqueta y sus atributos aparecen juntos sin espacio separador

Tabla V. Errores de codificación más frecuente en canales de sindicación de contenidos

Sin embargo, cada vez es más frecuente la incorporación de métodos de detección de errores en los sistemas de agregación de contenidos (Petrova-Antonova y Simov, 2011, p. 642). Por lo tanto, una forma efectiva de eliminar todos aquellos canales de sindicación no válidos es la importación directa de los enlaces en un agregador. Por regla general los canales de sindicación defectuosos, serán desestimados borrando automáticamente su enlace de la lista. Los canales de sindicación correctamente formados y validados serán importados y constarán de alguna información básica como el título y la descripción que en todo caso el agregador pueda recopilar. No obstante, la importación de los canales de sindicación no implica que estén disponibles para su explotación. Para efectuar investigaciones sobre la producción informativa de los canales de sindicación es necesario realizar una tarea de categorización previa que permita indicar aspectos clave como la procedencia, tipo de fuente, factor geográfico, dominio temático, valor de la fuente de información, idioma y prioridad de recuperación.

2.6. Caso práctico: canales de sindicación de los medios digitales de Portugal

Recientemente se ha completado un estudio de la producción informativa de los medios de comunicación digitales de Portugal que se publicará en la revista Transinformação de la Universidade Estadual de Campinas (Blázquez Ochando, 2017) usando el método descrito en el presente artículo. Las semillas de enlaces utilizadas para la extracción de los canales de sindicación de Portugal pueden ser consultadas y descargadas desde las siguientes referencias:

- Semilla de sitios web de prensa de Portugal. http://mblazquez.es/wp-content/uploads/semilla-prensa-pt.txt
- Semilla de sitios web de radios digital de Portugal. http://mblazquez.es/wp-content/uploads/semilla-radio-pt.txt
- Semilla de sitios web de televisión de Portugal. http://mblazquez.es/wp-content/uploads/semilla-television-pt.txt

Después de aplicar el método de extracción masiva de canales de sindicación, se obtuvieron los siguientes canales de sindicación:

- Canales de sindicación de prensa de Portugal. http://mblazquez.es/wp-content/uploads/feeds-prensa-pt.txt
- Canales de sindicación de radio de Portugal. http://mblazquez.es/wp-content/uploads/feeds-radio-pt.txt
- Canales de sindicación de televisión de Portugal. http://mblazquez.es/wp-content/uploads/feeds-television-pt.txt

De esta forma se tienen todos los canales de sindicación de la prensa (889 feeds), radio (208 feeds) y televisión (231 feeds) de Portugal, con los que se pudo realizar un estudio exhaustivo de la producción informativa del país.

3. Conclusiones

A pesar de que la sindicación de contenidos ha sido abordada desde el punto de vista de la comunicación informativa y documental, no se ha profundizado en los métodos de recopilación masiva de canales de sindicación en dominios y áreas de conocimiento específicas.

El desarrollo de estudios informétricos, de producción de información periodística, de opinión, tendencias o incluso de producción científica, según el objeto de estudio, depende de una mayor exhaustividad en las fuentes de información utilizadas. Esto significa no restringir las investigaciones a las fuentes conocidas y abrir el campo de estudio a nuevas fuentes que están por descubrir.

La aplicación del método expuesto ayuda a localizar nuevas fuentes de información con las que complementar los estudios de la información. Con ello se determinan distintas fases lógicas en las que el investigador plantea un contexto sobre el que construye un vocabulario o léxico normalizado, especializado en el área de conocimiento sujeto al análisis. Tomando como referencia la terminología organizada, se proponen estrate-

gias de consulta que la combinan usando operadores de consulta avanzada en buscadores para obtener resultados más pertinentes posibles. De los resultados obtenidos se extraen los enlaces que serán procesados por herramientas web crawler para extraer los enlaces de los canales de sindicación. Finalmente se detectan errores en los canales de sindicación y se completa la información clasificatoria y descriptiva que caracteriza su contenido.

Aunque no se proporciona un análisis de los agregadores que permiten realizar estudios informétricos a partir de los canales de sindicación, se puede adelantar que originalmente no fueron diseñados con ese objetivo. Es por ello que los investigadores no cuentan con herramientas adecuadas para aprovechar todas las posibilidades que ofrecen estas fuentes de información. De este problema, pueden deducirse futuras líneas de investigación que aborden el desarrollo de agregadores de contenidos diseñados para realizar estudios sobre la producción informativa, capaces de clasificar la información automáticamente, categorizar los canales de sindicación y obtener las relaciones inherentes entre los contenidos recuperados, descubriendo la influencia y el impacto de la información publicada..

Notas

(1) Selección de pluggins para los navegadores Google Chrome y Mozilla Firefox especializados en la copia de enlaces:
http://chrome.google.com/webstore/search/copy+links
http://addons.mozilla.org/firefox/search/?q=copy+links

(2) El buscador WauSearch permite exportar los resultados de las consultas en listas de enlaces de forma automática (http://www.wausearch.com/).

(3) Google2down es un servicio de exportación masiva de resultados de Google y Google Scholar (http://www.mblazquez.es/google2down/).

Referencias

Baeza-Yates, R.; Castillo, C.; Graells, E. (2008). Características de la web chilena 2007. // Technical Report, Center for Web Research, University of Chile.

Baporikar, M.; Salvi, S.; Sowany, V.; Sakhare, N. S. (2015). An approach towards news alert systems. // International Journal of Computer Science and Mobile Computing. 4:11 (noviembre 2015) 159-163.

Blázquez Ochando, M. (2013a). Sistemas de recuperación e internet: Metadescripción, procesamiento, webcrawling, técnicas de consulta avanzada, hacking documental y posicionamiento web. // Madrid: mblazquez.es

Blázquez Ochando, M. (2013b). Desarrollo tecnológico y documental del webcrawler Mbot: prueba de análisis web de la universidad española. // XIII Jornadas Españolas de Documentación Fesabid. (mayo 2013).

Blázquez Ochando, M. (2017). Método para el estudio de la producción informativa: Medios digitales de Portugal. // Transinformação. 29:1.

BuiltWith (2016). CMS technologies Web Usage Statistics. http://trends.builtwith.com/cms/ (2016-02-01).

Eito-Brun, R. (2015). Context-based aggregation of archival data: the role of authority records in the semantic landscape. // Archival Science. 15:3 (February 2014) 217-238.

Google. (2015). Google Web Search API (Deprecated). https://developers.google.com/web-search/docs/

Google. (2016). Google Alerts. https://www.google.es/alerts

Hammersley, B. (2003). Content syndication with RSS. // Sebastopol: O'Reilly, 2003.

Houssos, N.; Stamatis, K.; Banos, V.; Kapidakis, S.; Garoufallou, E.; Koulouris, A. (2011). Implementing enhanced OAI-PMH requirements for Europeana. // Research and Advanced Technology for Digital Libraries. Berlin: Springer, 2011.

Huang, Y.; Schuehle, J.; Porter, A. L.; Youtie, J. (2015). A systematic method to create search strategies for emerging technologies based on the Web of Science: illustrated for Big Data. // Scientometrics. 105:3 (July 2015) 2005-2022.

Hurtado, J. (2015). Automated System for Improving RSS Feeds Data Quality. // arXiv e-prints. http://arxiv.org/pdf/1504.01433v1 (2016-01-14).

Lee, B. S.; Im, J. W.; Hwang, B. Y.; Zhang, D. (2008). Design of an RSS Crawler with Adaptative Revisit Manager // SEKE. 219-222.

Lubbers, M. (2015). Adaptable Crawler Specification Generation System for Leisure Activity RSS Feeds. // Nijmegen: Radboud Universiteit, 2015.

McCown, F.; Nelson, M. L.; Van de Sompel, H. (2009). Everyone is a curator: human-assisted preservation for ore aggregations. // arXiv e-prints. (Consulta 2016-01-15) http://arxiv.org/pdf/0901.4571v1

Nieto, J. Y. (2015). Las revistas sociales personalizadas a través de agregadores compiten con el resto de medios informativos digitales. // Ambitos: Revista internacional de comunicación. 28 (July 2015) 5-13.

Peis, E.; Herrera-Viedma, E.; Morales-del-Castillo, J. M. (2008). Modelo de servicio semántico de difusión selectiva de información (DSI) para bibliotecas digitales. // El profesional de la información. 17:5 (diciembre 2007) 519-525.

Petrova-Antonova, D.; Simov, R. (2011). jQRSS: a jQuery plugin for RSS and Atom feeds parsing. Proceedings of the 12th International Conference on Computer Systems and Technologies. // ACM. (June 2011) 641-646.

Prient, C.; Lloret, E.; Palomar, M. (2012). Análisis de la calidad de la información recuperada por sistemas de alertas en el dominio Químico Textil. // II Congreso Español de Recuperación de Información CERI. (junio 2012).

Thatcher, A. (2008). Web search strategies: The influence of Web experience and task type. // Information Processing & Management. 44:3 (September 2007) 1308-1329.

Thelwall, M.; Prabowo, R.; Fairclough, R. (2006). Are raw RSS feeds suitable for broad issue scanning? A science concern case study. // Journal of the American Society for Information Science and Technology. 57:12 (agosto 2006) 1644-1654.

Viseur, R. (2012). Create a Specialized Search Engine – The Case of an RSS Search Engine // Proceedings of the International Conference on Data Technologies and Applications. 245-248.

Enviado: 2016-02-05. Segunda versión: 2017-03-28.
Aceptado: 2017-05-09.