Google scraping

MANUEL BLÁZQUEZ OCHANDO
Universidad Complutense de Madrid, España

INTRODUCCIÓN

El permanente crecimiento de la web y su indización en los principales buscadores ha propiciado un entorno documental de gran riqueza para los profesionales de la información. Debido a la magnitud y heterogeneidad del entorno, la extracción de contenidos relevantes se ha convertido en una tarea complicada, si no se disponen de las técnicas y métodos adecuados. Una solución a estos problemas es la *minería de datos* que consiste en la recolección sistemática de información procedente de documentos, páginas y sitios web estratégicos, que en origen podrían ser desconocidos para el investigador y que ayudan a formar nuevos conocimientos (Klösgen, W., y J. M. Zytkow, 2002).

Lamentablemente la minería de datos se basa en la asignación predeterminada de una serie de fuentes de información que serán el objetivo de los programas de rastreo. El investigador en todo caso condiciona la recopilación de datos con su conocimiento previo del entorno documental. Esto significa que sólo es posible recuperar una pequeña parte del todo a menos que el investigador indice Internet para discriminar qué fuentes de información especializadas son válidas y cuáles no. Esto origina diversas preguntas

¿Es posible obtener la información de un área de conocimiento en la web usando técnicas de minería de datos? ¿Puede el investigador conocer todas las fuentes de información usando programas de tipo *web crawler*? Como se ha explicado, la minería de datos prospecta una serie de recursos conocidos, pero difícilmente puede localizar recursos que el investigador ignora.

Para descubrir toda la información publicada sobre un área de conocimiento, sería necesario rastrear todas las direcciones IP registradas en la ICANN[1] e indizar sus contenidos para proceder a su filtrado posterior. Ésta tarea, del todo inasumible para la mayoría de investigadores,[2] la desarrollan los buscadores globales de la Web. Por tanto resulta lógico pensar que se necesitan programas que faciliten la recopilación de datos en éstos buscadores. Con este enfoque se vienen desarrollando los programas de tipo *SiteScraper* (Penman, Baldwin y Martinez, 2009) cuya técnica sería conocida con la denominación *web scraping* para aludir al proceso de raspado, análisis y extracción de datos de un sitio web mediante patrones.[3]

El tema de *web scraping sobre buscadores* no ha sido muy profuso en la literatura científica excepto en patentes como se explicará a continuación. Si se realiza la consulta (*scraping google* o *google scraping*) en el buscador académico *Google Scholar*, se obtienen únicamente 51 resultados de los que sólo 2 están directamente relacionados con la recolección masiva de datos de los buscadores. Uno de ellos corresponde al libro de (Calishain y

---

1 *Internet Corporation for Assigned Names and Numbers*. Organización encargada de la gestión del redireccionamiento de las *DNS Domain Name Server* a las direcciones IP de los servidores web conectados a Internet.

2 Aunque los investigadores dispongan de programas *web crawler* capaces de rastrear millones de sitios web, no pueden asumir sin los medios adecuados, el análisis de todo Internet. Se necesita una infraestructura muy desarrollada con centenares de servidores y programas *crawler* recopilando contenidos y almacenándolos en forma óptima para su posterior indización, clasificación y recuperación.

3 *XPath* es el método de selección de nodos en un objeto DOM que almacena el código fuente de una página web. Esto permite crear patrones reconocibles como etiquetas y atributos propios del lenguaje de marcado HTML.

Dornfest, 2003), titulado *Google Hacks*, que presenta el extinto API de Google para descargar sus páginas de resultados. La otra referencia corresponde al trabajo de (Ruecker y Devereux, 2004) en el que desarrollan un programa de *web scraping* para *Google News*.

Si se consulta (*web scraping*) como técnica aplicada a sitios y recursos en general, sin hacer mención a los buscadores, el número de resultados aumenta exponencialmente hasta las 1890 publicaciones científicas. Estos datos dan una idea de la importancia que tiene la técnica en el ámbito de las Ciencias de la Documentación. Algunas de las investigaciones más relevantes sobre *web scraping* tienen por objeto el desarrollo automático de la web semántica (Watson, 2009), la creación de webs sin intervención humana o autowebs (Thomsen *et al*., 2012), el desarrollo de modelos para compilar precios y comparar productos (Yibing, Zidong y Hongbo, 2014) o el análisis del discurso político de diversas fuentes de información (Hernández *et al*., 2015).

La literatura de patentes es muy destacada en torno al *web scraping* con más de 100 invenciones registradas sobre los procesos de descarga de contenidos de páginas web (Salerno y Boulware, 2006); la indización de textos por medio de scraping (Khan, 2012) y la protección de la información frente a esta técnica (Wetterström y Andersson, 2009).

Objeto de estudio

El objeto de la investigación es la experimentación de la técnica de *web scraping* en buscadores. En concreto se ha elegido el buscador Google por ser el más relevante y popular, así como su versión académica *Google Scholar*. Algunas de las razones que justifican su desarrollo son:

- *Web scraping sobre buscadores* es una forma de crear *big data* utilizando la principal fuente de información que indiza toda la web.

- Podrían crearse *scripts* que permitieran rastrear los contenidos de los buscadores y generar bases de conocimiento especializadas.
- Los contenidos recuperados se podrían clasificar más fácilmente al ser devueltos con *consultas compuestas* por las palabras clave o descriptores.
- Uniendo la técnica de *web crawler* a la de *web scraping* se podrían rastrear sectores de la web desconocidos para los investigadores.
- Los contenidos recuperados mediante *web scraping* permitirían desarrollar buscadores más especializados con un bajo costo de desarrollo.
- Podrían crearse nuevas utilidades de comparación basadas en los resultados de los buscadores.
- Pueden elaborarse nuevos estudios métricos y alt-métricos a partir de los resultados proporcionados por los buscadores.

La investigación parte de la premisa de que los buscadores son la llave de acceso a la información publicada en Internet. Por tanto, sería lógico ser capaces de recuperar la información devuelta por los buscadores ante una serie de palabras clave, descriptores o consultas dirigidas por los investigadores. En consecuencia ¿sería posible extraer toda la información necesaria de un buscador? ¿Qué dificultades se podrían encontrar? ¿Qué técnicas o métodos pueden emplearse para descargar la información? ¿Qué dilemas o cuestiones éticas y legales plantearía el aprovechamiento masivo de datos en torno a un buscador?

Metodología

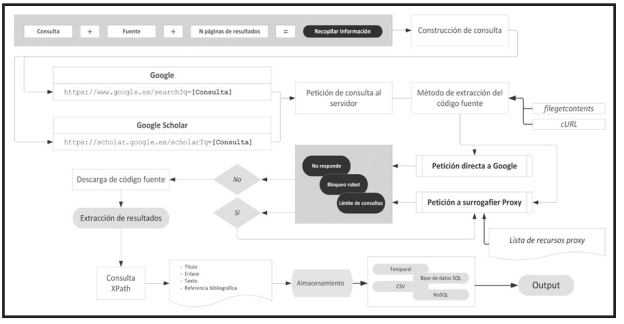
Una de las patentes con mayor impacto en el sector del *web scraping* (Nikovski y Esenther, 2008) lo define como el proceso de análisis de páginas web para la extracción de datos. El método comprende las siguientes etapas: 1) Descarga del código fuente de la página web para crear una plantilla DOM. Esto permite almacenar

temporalmente el código HTML de la página y estructurar su contenido mediante nodos, identificados como etiquetas embebidas y jerarquizadas. 2) La etapa de selección de los nodos o elementos de la plantilla DOM, como por ejemplo las etiquetas de título, enlace o párrafo que contienen la información demandada. Esto se lleva a cabo mediante expresiones *XPath* con las que se construyen las rutas y sub-rutas de consulta. 3) Proceso de almacenamiento de la información para su posterior aprovechamiento. Éste método también pudo ser comprobado en la investigación sobre *web spoofing* (Blázquez, 2015) para el estudio del comportamiento de usuario.

El método para desarrollar el experimento de *web scraping* con Google se basa en las mismas etapas. Sin embargo, existen otras consideraciones que deben ser incluidas en el proceso. 1) Definición de la dirección URL de consulta del buscador e identificación de sus variables. Esto es conocer todos los elementos que componen un enlace para indicar aspectos como las fechas extremas de consulta, paginación, idioma, país y operadores especiales. 2) Identificación de estructuras de datos que contienen la información que se pretende recuperar. Consiste en el análisis de las etiquetas contenedoras y sus rasgos caracterizadores que faciliten un acceso unívoco y directo. Por ejemplo, las etiquetas que poseen un atributo con una denominación única o un valor común a los contenidos que se pretenden recuperar. La inspección de estos aspectos favorece el proceso de recopilación mediante *XPath*, ya que permite construir mejores rutas de acceso a los datos. 3) Etapa de descarga del código fuente. Cuando se trata de un sitio web generalista, la descarga del código no representa ninguna dificultad. Esto no sucede en el caso de los buscadores, ya que presentan limitaciones para programas de scraping y robots. De hecho las respuestas más frecuentes son el bloqueo de la página de resultados, la presentación de pruebas anti-robot y el bloqueo por dirección IP. Si bien las contramedidas de los buscadores pueden detener a la mayoría de los programas de *web scraping*, se pueden añadir alternativas que evaden los controles del buscador. Se trata del uso de recursos proxy que permitan enmascarar la dirección IP real por otra. 4) Comprobación de la integridad del código fuente y

creación de la plantilla DOM. Como se indicó, este paso prepara la información contenida en el código para su recopilación. 5) Proceso de selección y extracción *XPath* mediante rutas de acceso a los elementos de la plantilla DOM. Se utilizan los rasgos diferenciadores que fueron identificados en el segundo paso. 6) Almacenamiento de los datos, tabulación y preparación para su representación. El método puede ser consultado en la *Figura 1*.

Figura 1. Proceso de web scraping en Google.



Fuente: <http://mblazquez.es/wp-content/uploads/google-scraping.png>

Programa de scraping de Google

El experimento de *web scraping* ha sido diseñado para descargar los contenidos más sustanciales, de las páginas de resultados de los buscadores Google y *Google Scholar*. Para ello ha sido creada una interfaz sencilla que está compuesta por una caja de consulta, un selector de buscador, un selector para activar la función de rastreo complementaria y otro para determinar el número de páginas de resultados sujetas a recopilación. El programa puede ser consultado para su verificación en la dirección <www.mblazquez.es/google2down>

El programa ha sido sometido a pruebas de estrés para determinar el punto de ruptura del servicio del buscador Google, antes de ser aplicado el bloqueo preceptivo. En la mayoría de los casos, las consultas sencillas han permitido un máximo de 100 peticiones al servidor. Se pudo comprobar que según se aumen-

ta la complejidad de las consultas con operadores *intitle*, *intext*, *inurl*, *filetype*, el buscador bloquea con mayor frecuencia las peticiones al servidor. Se puede afirmar que la sencillez de la consulta es directamente proporcional al número de peticiones que acepta Google. Cuanto más compleja sean las consultas, menos consultas por dirección IP permitirá el buscador. También implica mayor uso de los recursos de computación y resulta lógico que el acceso a consultas avanzadas automáticas sea más restringido. El bloqueo de los buscadores es resuelto mediante el enmascaramiento de la dirección IP y el uso de programas de tipo *proxy*. En este caso se usó el programa *Surrogafier* de *Brad Cable* (2015). Así cuando no se detectan resultados del buscador, el programa de *web scraping* transmite la petición de consulta al *proxy*, enmascarando el origen de la petición. Al no figurar la nueva dirección IP entre la lista de referencias bloqueadas, permite la recepción de la consulta y la consiguiente respuesta, descarga y extracción de sus datos. No obstante, el uso de un programa *proxy* no es garantía de funcionamiento permanente. De hecho Google puede detectar que las peticiones del *proxy* se desarrollan con una serie de direcciones IP específicas. Por ello el método óptimo consiste en disponer de una red de programas *proxy* que gestionen un rango importante de direcciones IP. De esta forma las direcciones comprometidas se desbloquean al cabo de 24 horas para volver a estar disponibles.

Figura 2. Impresión de pantalla del experimento de web scraping.

Fuente: Elaboración propia

La selección de elementos de Google y *Google Scholar* es posible construyendo las rutas *XPath* que se muestran en la *Tabla 1*. Pueden observarse algunos elementos distintivos como los atributos *class=st*, *id=gs_cit0*, i*d=gs_cit1*, *id=gs_cit2* que permiten recuperar la cita textual de los resultados, su referencia bibliográfica en formato APA, ISO y MLA.

Tabla 1.
Rutas XPath para la selección de resultados en Google y Google Scholar.

Web scraping en Google
• Título $xpath11->query("//h3/a[contains(@href,'/url')]")->item($i11)->nodeValue; • Enlace $xpath11->query("//h3/a[contains(@href,'/url')]/@href")->item($i11)->nodeValue; • Texto $xpath11->query("//span[@class='st']")->item($i11)->nodeValue;
Web scraping en Google Scholar
• Título $xpath11->query("//h3/a")->item($i11)->nodeValue; • Enlace $i11tem_link = $xpath11->query("//h3/a/@href")->item($i11)->nodeValue; • Referencia APA $xpath22->query("//div[@id='gs_cit0']")->item($i22)->nodeValue; • Referencia ISO $xpath22->query("//div[@id='gs_cit1']")->item($i22)->nodeValue; • Referencia MLA $xpath22->query("//div[@id='gs_cit2']")->item($i22)->nodeValue;

Fuente: Elaboración propia

Otro aspecto probado es la función de rastreo secundario de los resultados. Los enlaces recopilados son procesados mediante web crawler. En este caso se utiliza el código del programa *Mbot* (Blázquez, M. 2013). Este método amplía la información de los resultados obtenidos por el buscador, ya que permite rastrear todos

sus enlaces, recursos y texto completo. Esto significa que es posible crear nuevos buscadores basados en los ya existentes. En la *Figura 3* se observa cómo los resultados pueden ser analizados a voluntad por el usuario, obteniendo el texto indexable y los enlaces recuperados.

Figura 3.
Función de rastreo de resultados desde el programa de scraping.

Fuente: Elaboración propia

Conclusiones

El experimento demuestra que es posible usar técnicas de *web scraping* en buscadores y descargar la información que se necesite para componer una nueva base de conocimiento.

Las limitaciones del *web scraping* surgen cuando se desea recopilar cantidades masivas de datos. Los buscadores identifican el equipo que está realizando las descargas de contenidos y bloquean su trabajo mediante medidas de seguridad *captcha* que deben ser desbloqueadas manualmente, o bien, añadiendo la dirección IP del usuario en la lista negra de accesos no autorizados.

Las medidas de seguridad que presentan los buscadores son evitables, tal como queda demostrado en la investigación, siempre que se usen listas de servidores proxy que logren enmascarar o cambiar la dirección IP del usuario. De esta forma, cuando una dirección IP queda bloqueada, los programas de *scraping* pueden usar un *proxy* para continuar el trabajo de descarga.

La técnica de *web scraping* en el ámbito de la documentación permite abrir una nueva vía para la investigación. La creación de bases de conocimiento especializadas de forma automática, la recuperación y clasificación masiva de contenidos disponibles en la Web, la vigilancia informacional de alcance global, la creación de nuevos buscadores súper-especializados, la creación de *big data* en cualquier área de conocimiento y propósito son algunos ejemplos de aplicación que pueden esperarse del empleo de éstas técnicas.

Discusión, uso ético de la información: implicaciones y desafíos

Las aplicaciones del *web scraping* orientado a buscadores son muy prometedoras desde un punto de vista científico, tecnológico, social y económico, pero también es cierto que plantea diversos dilemas éticos. Por una parte el acceso a la información pública en la web es libre y resulta un argumento fuerte en favor de poder descargar los contenidos desde un buscador. Pero por otra parte, la tecnología y los medios necesarios para indexar y dar acceso a los miles de millones de sitios web, no son gratuitos. Son el resultado de un negocio basado en la recuperación de información y la publicidad. Por tanto cabe preguntarse ¿hasta qué punto podrían ser empleadas las técnicas de *web scraping* para generar nuevos productos basados en la recuperación de información?

Parece lógico que la aplicación científica del *web scraping* no debería perjudicar a los principales buscadores, ya que el propósito principal es la ampliación del conocimiento en beneficio de la sociedad y de la propia ciencia. Sin embargo si se diera el caso de copiar la información de los buscadores para competir con ellos bajo otra marca, se produciría un caso sin precedentes. Debe tenerse en cuenta que las direcciones de los sitios web son públicas y de libre acceso. Para acceder a ellas se utilizan los servicios de recuperación de un buscador de acceso abierto y cuyos resultados son descargados mediante *web scraping*. El resultado de las descargas masivas es la conformación de una nueva base de conocimiento para un nuevo producto o servicio. Para complicar aún más la

trama, puede afirmarse que copiar o referenciar los enlaces de las páginas de resultados de un buscador no es un acto delictivo por sí solo. Esto se debe a que es, en esencia, la misma tarea que desarrollan los buscadores cuando rastrean la web.

En relación a las cuestiones socio-económicas que podrían derivarse de la técnica de *web scraping*, un ejemplo claro corresponde a los buscadores de comparación de productos y servicios. Son programas especializados en la selección de sitios y páginas web de las que extraen información precisa con la que operan sus comparativas. Pero, qué ocurriría si se descargara un sector específico de la web a través de un buscador, para crear un producto de información, y posteriormente se comercializara ¿Sería reprobable el uso de buscadores para crear el producto? ¿Sería positivo por abrir nuevos mercados y crear puestos de trabajo?

La información publicada en Internet es cada vez más accesible como consecuencia del uso de buscadores y técnicas basadas en *crawling* y *scraping*. En la práctica puede afirmarse que los contenidos que no están protegidos en redes privadas son de acceso y dominio público. Esto plantea nuevos interrogantes y obliga a la comunidad científica a pensar en formas más sencillas y efectivas de proteger la información, pero también de abrir Internet al libre desarrollo de tecnologías de scraping como un subsector más de la industria.

El futuro del *web scraping* y el *big data* están directamente vinculados al profesional de la información y en particular a la figura del documentalista. Tareas como la selección de los datos, su procesamiento, almacenamiento, tabulación y análisis deberán ser dominadas para lograr una inteligencia competitiva más eficaz, crear nuevos servicios y productos de información que ayuden a dinamizar la economía y la sociedad.

Bibliografía

Blázquez, M. (2013). "Desarrollo tecnológico y documental del webcrawler Mbot: prueba de análisis Web de la universidad española". *XIII Jornadas Españolas de Documentación FESABID*. (23-25 mayo).

.——— (2015). "Técnica de suplantación de páginas web para el estudio de la usabilidad y experiencia de usuario: metodología y estudio de caso". *VII Encuentro EDICIC*. (16-17 noviembre).

Cable, B. (2015) *Surrogafier*. Disponible en: <https://github.com/BCable/surrogafier>

Calishain, T., y Dornfest, R. (2003). *Google Hacks*. O'Reilly Media.

Hernández, A. T., E. G. Vázquez, C. A. B. Rincón, J. M. García, A. C. Maldonado y R. I. Orozco. (2015) "Metodologías para análisis político utilizando Web Scraping". *Research in Computing Science*, vol. 95, 113-121.

Khan, S. (2012). *U.S. Patent Application No. 13/485,703*. Disponible en: <https://www.google.com/patents/US20120310914>

Klösgen, W. y J. M. Zytkow. (2002). *Handbook of Data Mining and Knowledge Discovery*. Oxford University Press, Inc.

Nikovski, D. N. y A. W. Esenther. (2008). *U.S. Patent Application No. 12/239,859*. Disponible en: <https://www.google.com/patents/US20100083095>

Penman, R. B., T. Baldwin y D. Martinez. (2009). *Web Scraping Made Simple with Sitescraper*.

Ruecker, S. y Z. Devereux. (2004). "Scraping Google and Blogstreet for Just-in-Time Text Analysis". *CaSTA-04*, *The Face of Text*, Ontario: McMaster University.

Salerno, J. J. y D. M. Boulware. (2006). *U.S. Patent No. 7,072,890*. Washington, DC: U.S. Patent and Trademark Office. Disponible en: <www.google.com/patents/US7072890>

Thomsen, J. G., E. Ernst, C. Brabrand y M. Schwartzbach. (2012). "Webself: A Web Scraping Framework". *Web Engineering*. Springer Berlin Heidelberg.

Watson, M. (2009). "Using Web Scraping to Create Semantic Relations. Scripting Intelligence: Web 3.0". *Information Gathering and Processing*, 205-228.

Wetterström, R. y S. Andersson. (2009). *U.S. Patent Application No. 13/000,157*. Disponible en: <www.google.es/patents/US20110185434>

Yibing, S., Z. Zidong, y L. Hongbo. (2014). "A Model of Compiling Price Index Based on the *Web Scraping* Technology". *Statistical Research*, vol. 10, núm. 15.