¿QUÉ ES EL WEB SCRAPING?

AUTOR: RICARDO CASTILLO BARQUERO

NOVIEMBRE: 2020

San Marcos

Introducción

Los motores de búsqueda, como Google, utilizan desde hace tiempo los denominados rastreadores web o crawlers, que exploran Internet en busca de términos definidos por el usuario. Los rastreadores son tipos especiales de bots, que visitan una página web tras otra para generar asociaciones con términos de búsqueda y categorizarlos. El primer rastreador web se creó ya en 1993, cuando se presentó el primer motor de búsqueda: Jumpstation.

Entre estas técnicas de rastreo se incluye el web scraping o webharvesting. En este artículo veremos cómo funciona, para qué se utiliza y cómo se puede bloquear en caso necesario.

Contenido

¿Qué es el web scraping?

Los motores de búsqueda, como Google, utilizan desde hace tiempo los denominados rastreadores web o crawlers, que exploran Internet en busca de términos definidos por el usuario. Los rastreadores son tipos especiales de bots, que visitan una página web tras otra para generar asociaciones con términos de búsqueda y categorizarlos. El primer rastreador web se creó ya en 1993, cuando se presentó el primer motor de búsqueda: Jumpstation.
Entre estas técnicas de rastreo se incluye el web scraping o webharvesting.

Web Scraping: definición

Durante el web scraping (del inglés scraping = arañar/raspar) se extraen y almacenan datos de páginas web para analizarlos o utilizarlos en otra parte. Por medio de este raspado web se almacenan diversos tipos de información: por ejemplo, datos de contacto, tales como direcciones de correo electrónico o números de teléfono, o también términos de búsqueda o URL. Estos se almacenan en bases de datos locales o tablas.
Con el web scraping se leen textos de páginas web para obtener información y almacenarla, de forma comparable al proceso automático de copiado y pegado. En el caso de la búsqueda de imágenes, el proceso se denomina acertadamente image scraping.

¿Cómo funciona el web scraping?

Dentro del scraping hay diferentes modos de funcionamiento, aunque en general se diferencia entre el scraping automático y el manual. El scraping manual define el copiado y pegado manual de información y datos, como quien recorta y guarda artículos de periódico y solo se lleva a cabo si se desea encontrar y almacenar alguna información concreta. Es un proceso muy laborioso que raras veces se aplica a grandes cantidades de datos.
En el caso del scraping automático, se recurre a un software o un algoritmo que analiza diferentes páginas web para extraer información. Se utiliza software especializado según el tipo de página web y el contenido.

Dentro del scraping automático, se diferencian varios modos de proceder:

- Analizador sintáctico: los analizadores sintácticos (o parsers) se utilizan para convertir un texto en una nueva estructura. Por ejemplo, en los análisis de HTML, el software lee un documento HTML y almacena la información. Un analizador DOM utiliza la representación de contenidos del lado del cliente en el navegador para extraer datos.
- Bots: un bot es un software dedicado a realizar determinadas tareas y automatizarlas. En el caso del web harvesting, los bots se utilizan para examinar páginas web automáticamente y recopilar datos.
- Texto: aquellos que tienen experiencia con la línea de comandos pueden aprovechar la función grep de Unix para buscar en la web determinados términos en Python o Perl. Este es un método muy sencillo para extraer datos, aunque requiere más trabajo que lautilización de un software.

¿Con qué fin se utiliza el web scraping?

El web scraping se utiliza para una gran variedad de tareas, por ejemplo, para recopilar datos de contacto o

información especial con gran rapidez. En el ámbito profesional, el scraping se utiliza a menudo para obtener ventajas respecto a la competencia. De esta forma, por medio del harvesting de datos, una empresa puede examinar todos los productos de un competidor y compararlos con los propios. El web scraping también resulta valioso en relación con los datos financieros: es posible leer datos desde un sitio web externo, organizarlos en forma de tabla y después analizarlos y procesarlos.
Google es un buen ejemplo de web scraping. El buscador utiliza esta tecnología para mostrar información meteorológica o comparaciones de precios de hoteles y vuelos. Muchos de los portales actuales de comparación de precios también utilizan el scraping para representar información de diferentes sitios web y proveedores.

¿Es legal el web scraping?

El scraping no siempre es legal. En primer lugar, los scrapers deben tener en cuenta los derechos de propiedad intelectual de los sitios web. El web scraping tiene consecuencias muy negativas para algunas tiendas online y proveedores, por ejemplo, si el posicionamiento de su página se ve afectado debido a los agregadores. No es raro, por lo tanto, que una empresa se querelle contra un portal de comparaciones para impedir el web scraping. En uno de estos casos, el Tribunal Superior Regional de Fráncfort dictaminó en 2009 que una compañía aérea debía permitir el scraping por parte de portales comparativos porque, al fin y al cabo, su información es de libre acceso. No obstante, la aerolínea tenía la posibilidad de recurrir a medidas técnicas para evitarlo.
El scraping es legal, por lo tanto, siempre y cuando los datos recabados estén disponibles libremente para terceros en la web. Para garantizar la legalidad del web scraping, hay que tener en consideración lo siguiente:

- Observar y cumplir con los derechos de propiedad intelectual. Si los datos están protegidos por estos derechos, no se pueden publicar en ninguna otra parte.
- Los operadores de las páginas tienen derecho a recurrir a procesos técnicos para evitar el web scraping que no pueden ser eludidos.
- Si, para la utilización de datos, se requiere el registro de usuarios o un contrato de utilización, estos datos no podrán ser aprovechados mediante scraping.
- No se permite la ocultación de publicidad, de términos y condiciones, ni de descargos de responsabilidad mediante tecnologías de scraping.

Aunque el scraping está permitido en muchos casos, puede utilizarse con fines destructivos o incluso ilegales. Por ejemplo, esta tecnología se utiliza a menudo para enviar spam. Los emisores pueden aprovecharla para, por ejemplo, acumular direcciones de correo electrónico y enviar mensajes de spam a estos destinatarios.

¿Cómo se puede bloquear el web scraping?

Para bloquear el scraping, los operadores de sitios web pueden adoptar diferentes medidas. Por ejemplo, el archivo robots.txt se utiliza para bloquear bots de buscadores. Por lo tanto, el scraping automático también se evita por medio de bots de software. Asimismo, es posible bloquear direcciones IP de bots. Los datos de contacto e información personal se pueden ocultar de manera selectiva: los datos sensibles, tales como los números de teléfono, se pueden proporcionar en forma de imagen o como CSS, lo que dificulta el scraping de los datos. Además, existen numerosos proveedores de servicios antibotde pago que pueden establecer un firewall. Con Google Search Console también es posible configurar notificaciones para informar a los operadores de sitios web en caso de que sus datos se utilicen para scraping.

Conclusiones y recomendaciones

El web scraping se utiliza para una gran variedad de tareas, por ejemplo, para recopilar datos de contacto o información especial con gran rapidez. En el ámbito profesional, el scraping se utiliza a menudo para obtener ventajas respecto a la competencia. De esta forma, por medio del harvesting de datos, una empresa puede examinar todos los productos de un competidor y compararlos con los propios. El web scraping también resulta valioso en relación con los datos financieros: es posible leer datos desde un sitio web externo, organizarlos en forma de tabla y después analizarlos y procesarlos.

Google es un buen ejemplo de web scraping. El buscador utiliza esta tecnología para mostrar información meteorológica o comparaciones de precios de hoteles y vuelos. Muchos de los portales actuales de comparación de precios también utilizan el scraping para representar información de diferentes sitios web y proveedores. Por otra parte, existen numerosos proveedores de servicios antibot de pago que pueden establecer un firewall. Con Google Search Console también es posible configurar notificaciones para informar a los operadores de sitios web en caso de que sus datos se utilicen para scraping.

Referencias bibliográficas

- Desarrollo Web. (2020). ¿Qué es el web scraping? Recuperado de https://www.ionos.es/digitalguide/paginas-web/desarrollo-web/que-es-el-web-scraping/

San Marcos

www.usanmarcos.ac.cr

San José, Costa Rica