Portada Portal de la comunidad Actualidad Cambios recientes Páginas nuevas Página aleatoria Ayuda Donaciones Notificar un error

Imprimir/exportar Crear un libro Descargar como PDF

Herramientas

Lo que enlaza aquí Cambios en enlazadas

Subir archivo

Versión para imprimir

Páginas especiales Enlace permanente Información de la página

Wikidata Citar esta página

O

Elemento de

En otros idiomas العربية Català Deutsch English Français Islenska Italiano 日本語 Latviešu Nederlands

Cрпски / srpski

Türkçe

Українська

Artículo Discusión

Leer Editar Ver historial

Buscar

⊗

Q

Web scraping



Este artículo o sección necesita referencias que aparezcan en una publicación acreditada, como revistas especializadas, monografías, prensa diaria o páginas de Internet fidedignas. Este aviso fue puesto el 1 de junio de 2014. Puedes añadirlas o avisar al autor principal del artículo en su página de discusión pegando: {{subst:Aviso referencias|Web scraping}} ~~~~

¡Participa del concurso de edición La Mujer que nunca conociste! Crea y edita biografías de mujeres.

Web scraping es una técnica utilizada mediante programas de software para extraer información de sitios web. Usualmente, estos programas simulan la navegación de un humano en la World Wide Web ya sea utilizando el protocolo HTTP manualmente, o incrustando un navegador en una aplicación.

El web scraping está muy relacionado con la indexación de la web, la cual indexa la información de la web utilizando un robot y es una técnica universal adoptada por la mayoría de los motores de búsqueda. Sin embargo, el web scraping se enfoca más en la transformación de datos sin estructura en la web (como el formato HTML) en datos estructurados que pueden ser almacenados y analizados en una base de datos central, en una hoja de cálculo o en alguna otra fuente de almacenamiento. El término web scraping también está relacionado con la automatización de tareas en la Web, la cual simula la navegación de un humano utilizando un software de computadora. Alguno de los usos del web scraping son la comparación de precios en tiendas, la monitorización de datos relacionados con el clima de cierta región, la detección de cambios en sitios webs y la integración de datos en sitios webs. También es utilizado para obtener información relevante de un sitio a través de los rich snippets.

Indice [ocultar]

- 1 Técnicas
- 2 Cuestiones legales
- 3 Medidas para detener a los scrapers
- 4 Herramientas notables
- 5 Véase también
- 6 Tutoriales en la web

Técnicas [editar]

Web scraping es el proceso de recopilar información de forma automática de la Web. Es un campo con desarrollos activos, compartiendo un propósito en común con la visión de la Web semántica. Utiliza soluciones prácticas basadas en tecnologías existentes que son comúnmente ad hoc. Existen distintos niveles de automatización que las existentes tecnologías de Web Scraping pueden brindar:

- «Copiar y pegar» humano: algunas veces incluso las mejores técnicas de web scraping no pueden reemplazar la examinación manual de un humano, y a veces esta puede ser la única vía de solución cuando el sitio que tenemos en mente pone ciertas barreras para prevenir que se creen softwares para realizar tareas automáticas en este.
- Uso de expresiones regulares: una posible vía para extraer información de páginas webs pueden ser las expresiones regulares, aunque comúnmente no se recomienda. utilizarlas para parsear el formato HTML.
- Protocolo HTTP: páginas webs estáticas y dinámicas pueden ser obtenidas haciendo peticiones HTTP al servidor remoto utilizando sockets, etc.
- Algoritmos de minería de datos: muchos sitios webs tienen grandes colecciones de páginas generadas dinámicamente a partir de una base de datos. Datos de la misma categoría aparecen usualmente en páginas similares mediante un script o una plantilla. En la minería de datos, un programa detecta estas plantillas en un contexto específico y extrae su contenido.
- Parsers de HTML: Algunos lenguajes, como XQuery y HTQL pueden ser utilizados para parsear documentos, recuperar y transformar el contenido de documentos HTML.
- Aplicaciones para web scraping: existen muchas aplicaciones disponibles que pueden ser utilizadas para personalizar soluciones de Web Scraping. Estas aplicaciones pudieran reconocer automáticamente la estructura de cierta página o brindar una interfaz al usuario donde este pudiera seleccionar los campos que son de interés dentro del documento. De esta forma no es necesario escribir manualmente código para realizar estas tareas.
- Reconocimiento de información semántica: las páginas que son analizadas podrían incluir metadatos o cierta información semántica como anotaciones o comentarios, los cuales pueden ser usados comúnmente. Si estas anotaciones están en las mismas páginas, como sucede con los microformatos, estas podrían ser de utilidad cuando parseamos el DOM del documento. En otro caso, las anotaciones, organizadas en una capa semántica, son almacenadas y manejadas de forma separada desde otras páginas, por lo que los scrapers pueden recuperar estos esquemas y las instrucciones desde esta capa antes de analizar los documentos.

Cuestiones legales [editar]

El web scraping pudiera ir en contra de los términos de uso de algunos sitios webs. El cumplimiento de estos términos no está totalmente claro. Mientras que la duplicación de expresiones originales puede ser en muchos casos ilegal, en Estados Unidos la corte dictó en el caso Feist Publications v. Rural Telephone Service que la duplicación de hechos es permitida. Las cortes de Estados Unidos en ciertas ocasiones han reconocido que ciertos usos de los scrapers no deberían estar permitidos. Podría considerarse una computadora como una propiedad personal, y de esta forma el scraper estaría entrando sin autorización en esta propiedad. En el caso más conocido, eBay vs Bidder's Edge, la segunda empresa tuvo que parar de realizar peticiones automáticas al sitio de eBay. En este caso, Bidder's Edge pujaba automáticamente por ciertos productos en este sitio.

Uno de las principales pruebas de scraping involucró a American Airlines y a una empresa llamada FareChase. American Airlines ganó esta batalla, haciendo que FareChase parara de vender un software que le permitía a los usuarios comparar tarifas en línea si el sitio de American Airlines era incluido. La aerolínea dijo que las búsquedas de FareChase entraban sin autorización en los servidores cuando recopilaban la información públicamente disponible.

Southwest Airlines también ha sido víctima de prácticas de web scraping.

Aunque las decisiones actualmente tomadas no son uniformes, es difícil ignorar que un patrón está emergiendo, en el cual podemos ver que las cortes están preparándose para proteger el contenido propietario en sitios webs comerciales, previendo de esta forma que este sea utilizado sin el consentimiento de los propietarios de los sitios. Sin embargo, el grado de protección de estos contenidos aún no está establecido, y dependerá de el tipo de acceso realizado por los scrapers, de la cantidad de información recopilada y de el grado en el que afecten estos factores al propietario del sitio web.

Medidas para detener a los scrapers [editar]

El administrador de un sitio web puede utilizar varias técnicas para detener o disminuir los pedidos de los scrapers. Algunas técnicas incluyen:

- Añadir entradas al fichero robots.txt. Google y otros bots pueden ser detenidos de esta forma.
- Bloquear la dirección IP. Esto también bloqueará todos los accesos desde esa misma IP, por lo que los usuarios no podrán navegar por el sitio web si acceden desde ésta.
- Deshabilitar cualquier interfaz de programación de aplicaciones que el sitio web pudiera estar brindando.
- Los bots o scrapers algunas veces declaran quienes son, y gracias a esto pueden ser bloqueados. «googlebot» es un ejemplo. Algunos scrapers no hacen distinción entre ellos y un navegador común.
- Monitorear el exceso de tráfico proveniente de cierta IP.
- Añadir un captcha u otro sistema de verificación manual al sitio web. No se garantiza el completo bloqueo de los scrapers, pero mediante esta técnica se dificulta el acceso de los mismos a los sitios webs.
- Servicios comerciales antibots: algunas empresas ofrecen servicios antibots y antiscraping.
- Incrementar el uso de JavaScript y AJAX. De esta forma es más difícil para los scrapers simular las peticiones como si fueran un navegador común.

Herramientas notables [editar]

- Apache Camel
- Automation Anywhere Convertigo
- cURL
- Data Toolbar
- Firebug Greasemonkey
- HtmlUnit
- Node.js

- HTTrack
- Aptana Jaxer
- nokogiri

iMacros

- watir
- Wget WSO2 Mashup Server
- HtmlAgilityPack
- BeautifulSoup
- Scrapy

Véase también [editar]

- Minería de datos
- Mashup (aplicación web híbrida)
- Spamdexing Corpus lingüístico
- Araña web Metadato

Tutoriales en la web [editar]

Scraping Fácil Con Python

Categorías: World Wide Web | Minería de datos | Web semántica | Web 2.0 | Tecnología web

Esta página fue modificada por última vez el 29 ene 2016 a las 11:28.

El texto está disponible bajo la Licencia Creative Commons Atribución Compartir Igual 3.0; podrían ser aplicables cláusulas adicionales. Léanse los términos de uso para más información. Wikipedia® es una marca registrada de la Fundación Wikimedia, Inc., una organización sin ánimo de lucro







