Importancia de Web Scrapping Y Web Crawler En La Minería De Datos

Onofre Soto Ana Laura

ana_onofre@outlook.es

De la Cruz Sosa Carlos

cdelacruzupiita@gmail.com

Santiago Godoy Rafael

Unidad Profesional Interdisciplinaria en Ingeniería y Tecnologías Avanzadas, UPIITA

Resumen

Hoy en día la información está al alcance de quien posee internet y un dispositivo inteligente, ya que se puede saber desde la hora y fecha actual, hasta que sucedió tiempo atras del otro lado del mundo en tan solo unos instantes de tiempo, pero alguna vez nos hemos preguntado ¿Cómo es posible esto?, ¿Qué algoritmos se utiliza en los famosos buscadores en línea para que al realizar una búsqueda, estos encuentren y muestren páginas con contenido que nos pudiera ser de utilidad?, pues bien, desde principios de los noventas existe algo llamado web crawler y web scrapping, los cuales en el presente artículo se abordaran para mostrar como estos algoritmos que se utilizan para buscar y recolectar información de la red (crawler) para posteriormente procesarla (scrapping) y de esta forma lograr optimizar la búsquedas, lo cual es de gran ayuda cuando se desea encontrar información muy puntual sobre un tema, ya que al tener una inmensa cantidad de páginas en la red, es poco práctico estar observando cada página para poder resolver una consulta, puesto que tener tanta información en un mismo lugar no sirve de nada si no se le da una buena interpretación.

I. Introducción

Hoy en día vivimos en la era de la información, el mundo está interconectado gracias al uso masivo del internet, con lo cual conocer sobre cualquier tema resulta posible en cualquier lugar y momento para cualquier persona, solo es necesario tener un dispositivo inteligente con acceso a internet para realizar una búsqueda y acceder a varios de los miles de sitios que existen en la red, y es que como bien se dice, la información es poder, por lo que muchas empresas e instituciones invierten dinero y esfuerzo en la minería de datos, la cual según Ester Ribas[1] es “un conjunto de técnicas y tecnologías que permiten explorar grandes bases de datos, de manera automática o semiautomática, con el objetivo de encontrar patrones

repetitivos que expliquen el comportamiento de estos datos" , ya que a pesar de que hoy en día se tiene un gran cúmulo de información en un mismo lugar, no sirve de nada si no se le puede dar una correcta interpretación. Por este motivo es que se han desarrollado varios algoritmos a nivel de software para poder ayudar a la minería de datos a encontrar de forma más rápida y eficiente sitios donde haya información que pudiera ser de utilidad, dentro de los cuales se encuentran el web scrapping y el web crawler, los cuales abordaremos en este artículo.

II. Comencemos a recompilar datos

Las técnicas más usadas para recompilar información en la red son el web crawling y el web scrapping. El web crawler o también conocido en español como 'araña web' es un robot de la web que rastrea de forma sistematizada un sitio web para recoger información de las páginas que se encuentran ahí, generalmente al principio se tiene un semillero con ciertas URL´s conocidas, las cuales las arañas de la web (crawler) recorren para encontrar nuevas URL´s y así descargarlas al equipo local y añadirlas a la lista, de esta manera se logra tener acceso a diversas páginas, aunque nunca al 100% de los sitios existentes en Internet [2](Figura 1). Cabe mencionar que el crawler no solo sirve para motores de búsqueda, sino también para realizar estudios estadísticos de diversos tipos, crear eventos, verificar que los enlaces a los que se apunta estén activos, enviar e-mails cuando se cumpla ciertas condiciones entre otras cosas. Debido a que se puede hacer validaciones sobre el código HTML, y diversos parámetros que hay en un sitio web.

Figura 1. Funcionamiento de web Crawler.

En la minería de datos el crawler resulta bastante útil, ya que de acuerdo a la revista Muy interesante el proceso de esta tiene normalmente cuatro etapas principales:

1. Determinación de los objetivos
2. Procesamiento de los datos
3. Determinación del modelo
4. Análisis de los resultados

En la determinación de los objetivos se especifica que información se quiere conseguir, para que en el procesamiento de los datos se elija, filtre, y sintetice todos los datos que se tienen, en esta segunda etapa aparece el uso del crawler para la recolección de toda la información útil que sea posible. Una vez que esta etapa se concluye procedemos a la determinación del modelo, en donde dependiendo de la problemática que se tenga, se seleccionara un algoritmo en particular que nos permita darle solución, en las mayorías de la veces el web scrapping resulta de gran ayuda, ya que permite resolver los 5 mayores conflictos de esta tercera etapa, los cuales son: organización, regresión, fragmentación, unión y síntesis de los datos[3].

El web scraping es una solución tecnológica que extrae datos de sitios web de forma rápida, eficiente y automatizada, ofreciendo datos de una manera más sencilla de usar [4] (Figura 2). Existen diversos lenguajes de programación para desarrollar este tipo de herramientas como java, node y phyton. Hoy en día la más usada y con mayor auge es Python, puesto que la documentación de web scrapping en este lenguaje de programación tiene mayor documentación por su facilidad de programación en comparación con otros lenguajes de alto nivel como es Java.

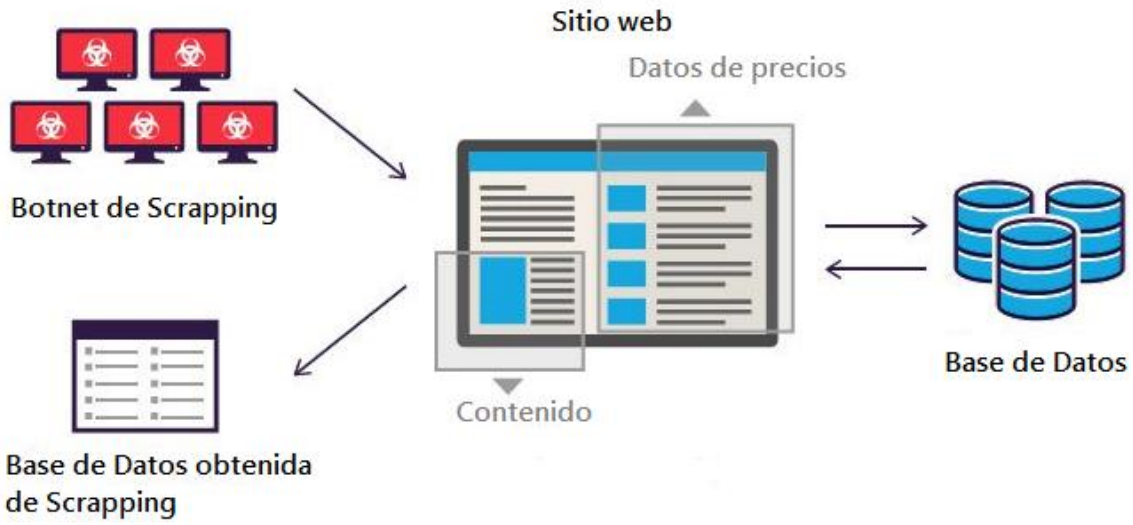

Figura 2. Funcionamiento de web Scrapping.

La última etapa de la minería de datos es el de análisis de los resultados, los cuales se obtienen con base en la información reunida en las fases anteriores. En la actualidad la minería de datos está en pleno auge ya que se hace uso de ella en diversas áreas comerciales como marketing, de la misma forma que en los sectores públicos de salud y seguridad, sin olvidar su empleo en las áreas de ciencia y tecnología, de donde tenemos claro ejemplo las búsquedas online, el procesamiento de lenguaje natural e inteligencia artificial [5].

II. El web Crawler y web Scrapping en acción

Es fácil ver que para poder realizar tareas de Scraping primero se tiene que hacer uso del web crawler, ya que primero se debe recuperar información, para después analizarla de manera más detallada y así extraer información más puntual.

Para comprender mejor esto, vamos a plantear dos situaciones diferentes: Supongamos que se desea conocer acerca de cuáles son los destinos turísticos más económicos en una determinada zona, como puede ser la ciudad de México, para lo cual tenemos 3 sitios web que tratan sobre 3 lugares turísticos en la ciudad, con ayuda del web Crawler lo que vamos hacer es que todos los enlaces que existan en estas 3 paginas, se van a guardar en una lista y se descargaran en el equipo local, de esta forma tendremos más sitios que hablen acerca de más de 3 destinos turísticos en la ciudad de México o relacionados con este tema, posteriormente al analizar estos sitios, implementando web Scrapping, se podrá extraer el precio de cada destino turístico y por ende obtener una búsqueda más precisa sobre lo que se desea.

Para el primer caso se planteó un problema meramente comercial, pero en este segundo caso observaremos que el Crawler y el Scraping podría ayudar al sector salud ya que plantearemos la búsqueda de las estadísticas sobre la incidencia del cáncer en la población de diversos países, para posteriormente vincularlo con información recompilada por instituciones oficiales y de prestigio que hablen de los factores que influyen para que una persona presente esta enfermedad, aunado a un análisis sobre la cultura, alimentación, geografía y costumbres de cada país, de esta manera se podrá observar patrones que nos indiquen que poblaciones son más propensas a padecer cáncer, debido a su entorno para que se puedan tomar medidas adecuadas en el sector salud de cada país.

Como se puede observar en los dos ejemplos planteados el crawler y scrapping son herramientas bastante útiles para la minería de datos, ya que prácticamente se logra obtener información muy puntual de forma automática, y sistematizada en poco tiempo, sin olvidar que el análisis de la información se vuelve más sencillo, dándonos la ventaja de tomar mejores decisiones en cualquier rubro donde se está aplicando la minería de datos.

III. Conclusiones

La sociedad en la que vivimos está llena de información que parece imposible que estemos desinformados, pero la verdad es que pocas personas saben cómo manejar la información a su favor, ya que al ver tantos datos en un mismo lugar, resulta casi imposible distinguir una cosa de otra, para lo que es importante que se implemente minería de datos la cual haga uso de web crawler y web scrapping, ya que con ayuda de ellos, podremos aprovechar la gran ventaja de vivir en un mundo interconectado, puesto que la información solo es una arma poderosa cuando se le da una buena interpretación y uso. Por ello es que el desarrollo de software basado en algoritmos que permitan mejorar y agilizar aún más las búsquedas en una red, son verdaderas joyas para cualquier persona, empresa o institución por todas las ventajas que tiene, pues permite ahorrar tiempo, dinero y recursos, que pueden ser requeridos en otras áreas.

Referencias

1. Ester Ribas. (18/ENE/2018). ¿Qué es el Data Mining o minería de datos?. 25/JUN/2018, de IEBS Sitio web: https://www.iebschool.com/blog/data-mining-mineria-datos-big-data/

2. Alexander Escobar. (12/AGO/2016). Web Crawler ¿Qué son? y ¿Cómo Funcionan?. 22/JUN/2018, de VUXMI Sitio web: http://vuxmi.com/web-crawler-que-son-y-como-funcionan/amp/

3. Anónimo (2014). ¿Qué es la minería de datos? De Muy Interesante. Recuperado 20/JUN/2018, Sitio web: https://www.muyinteresante.es/tecnologia/preguntas-respuestas/que-es-la-mineria-de-datos-311477406441

4. Osmar Castrillo-Fernández. (Deciembre 2015). Web Scraping: Applications and Tools. 22/JUN/2018, de European Public Sector Information Platform Sitio web: https://www.europeandataportal.eu/sites/default/files/2015_web_scraping_applications_and_tools.pdf

Riquelme, José C., Ruiz, Roberto, Gilbert, Karina, Minería de Datos: Conceptos y Tendencias. Inteligencia Artificial. Revista Iberoamericana de Inteligencia Artificial [en linea] 2006, 10 (primavera) : [Fecha de consulta: 25 de junio de 2018] Disponible en:<http://www.redalyc.org/articulo.oa?id=92502902