RECyT
Año 21 / Nº 32 / 2019 / 28–32

Técnica de extracción, transformación y carga de datos de estaciones meteorológicas

Extraction, transformation and loading technique of meteorological stations data

Héctor Ramón López[1,*]

1- Facultad de Ciencias Exactas, Físicas y Naturales, Universidad Nacional de San Juan, Av. José Ignacio de la Roza Oeste 590, San Juan, Argentina

* E-mail: lepezhr@yahoo.com.ar

Resumen

Climatología y Meteorología son ciencias que utilizan datos como la temperatura del aire, la humedad, el viento o la precipitación para tomar medidas; es esencial tener la cantidad más grande de datos para reducir el margen de error. Estos datos son adquiridos por estaciones meteorológicas automáticas que pertenecen a organismos públicos y en menor medida privados. La mayoría de los datos se publican en Internet, sin embargo, los datos requeridos no siempre están disponibles en formatos estándar para intercambio de información. El objetivo de este trabajo de investigación es la construcción de un algoritmo para la adquisición de datos meteorológicos, basado en una técnica conocida en inglés como web scraping. El proceso consiste en la extracción, normalizar los datos y luego almacenarlos, para ser utilizados por herramientas geográficas o análisis estadísticos.

Palabras clave: Meteorología; Datos; Algoritmo; Adquisición; Scraping.

Abstract

Climatology and Meteorology are sciences that use data such as air temperature, humidity, wind or precipitation to take measurements; it is essential to have the largest amount of data to reduce the margin of error. These data are acquired by automatic meteorological stations that belong to public and to a lesser extent private organizations. Most data is published on the Internet, however, the required data is not always available in standard formats for information exchange. The objective of this research work is the construction of an algorithm for the acquisition of meteorological data, based on a technique known in English as web scraping. The process consists in the extraction, normalizing the data and then storing them, to be used by geographic tools or statistical analysis.

Keywords: Meteorology; Data; Algorithm; Acquisition; Scraping

Introducción

La Climatología y Meteorología son ciencias que utilizan datos como la temperatura del aire, humedad, presión atmosférica, viento o precipitaciones; variables que se obtienen de estaciones de medición automáticas también llamadas por sus iniciales (E.M.A.). Las E.M.A. pertenecen a organismos públicos y en menor medida a privados. Sin embargo, a menudo los datos requeridos no están disponibles en formatos de intercambio estándar para su inmediata utilización. Una de las dificultades a la hora de realizar pronósticos, sistemas de alertas climáticos o reportes, radica en poder recopilar la mayor cantidad de datos dispersos por la gran red de redes, analizarlos en conjunto, y obtener conclusiones que sirvan para el trabajo de científicos, analistas o expertos.

El objetivo es la construcción de un algoritmo de selección o recolección de datos meteorológicos por la web, este proceso se conoce en inglés como técnica de scraping. Se puede conceptuar scraping o web scraping, como "un proceso que implica la recuperación de un documento de internet, generalmente de una página web en un lenguaje marcado, y el análisis de ese documento con el fin de extraer datos específicos para su uso en otro contexto" Turland, M. (2010)[1]. Cuando Turland se refiere a lenguaje marcado apunta al conjunto de etiquetas que contienen información adicional acerca de la estructura del texto o su presentación.

El proceso de recolección de datos se basa en acceder a fuentes de distinto origen, normalización de los datos adquiridos y su posterior almacenamiento, para luego poder ser utilizados por herramientas geográficas o específicas de análisis estadístico.

En este trabajo de investigación se desarrolló la técnica de extracción, transformación y carga requerida para la toma de datos en tiempo real de los distintos sitios web de las estaciones meteorológicas automáticas (E.M.A.) Además de desarrollar un sistema de prueba y monitoreo de los datos capturados.

Métodos

Un programa que inspecciona las páginas del World Wide Web (WWW) de forma metódica y automatizada y que consisten en una búsqueda y extracción se denominan web crawlers, Penman, R. B., Baldwin, T., & Martinez, D. (2009)[2]. Una de las formas más utilizadas se basa en crear una copia de todas las páginas web visitadas para su procesamiento por un motor que indexa las páginas proporcionando búsquedas rápidas.

Una variante al crawler es el scrapers, ver figura 1, cuya principal diferencia es la búsqueda de cierto tipo de información. Dentro de las múltiples formas de aplicar scraping, se utiliza la búsqueda de palabras claves, también denominado concordancia de expresiones regulares, que busca palabras coincidentes en la programación del HyperText Markup Language (HTML) del sitio web.

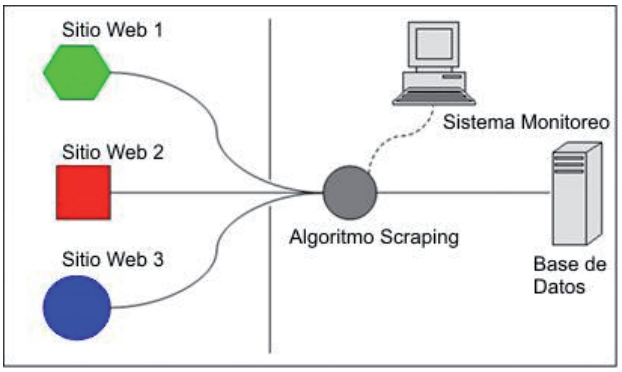

Figura 1: Representación del proceso de scraping.

La búsqueda de palabras claves, le permite al algoritmo de extracción obtener los datos de las estaciones meteorológicas, como por ejemplo la temperatura o la presión atmosférica. Para el desarrollo del algoritmo se utiliza código de programación Hypertext Preprocessor (PHP) que permite, a través de sus librerías, utilizar funciones para leer los datos de los sitios web de las distintas E.M.A. El proceso implica tener un diccionario de búsqueda y un patrón de sitios web definido. (Mehlführer, A. 2009) [3], es decir por ejemplo todas las estaciones meteorológicas de un organismo deberían tener un patrón definido para publicar los datos en un formato igual o similar, que simplifique la tarea del algoritmo de scraping.

Para la transformación de los datos es necesario convertir las variables que lo requieran en base a cálculos matemáticos. Por ejemplo, si se extrae un dato temperatura de un sitio web en grado Fahrenheit, se debe convertir a Celsius si así lo requiere.

Lenguaje de desarrollo del algoritmo

El lenguaje utilizado para el desarrollo del algoritmo es Hypertext Preprocessor (PHP), de uso general interpretado por el servidor, originalmente diseñado para el desarrollo web de contenido dinámico.

PHP es práctico para el desarrollo de aplicativos ya que permite interactuar a través de sus drivers de conexión con diversos motores de bases de datos o archivos de textos. Además ofrece librerías que permite implementar los procesos de extracción de datos, estas funciones se encuentran contenidas en la extensión (cURL).

Extensión cURL

Este conjunto de funciones han sido añadidas a PHP en su versión 4.0.2. cURL, https://curl.haxx.se/, es una librería para conectar con servidores y poder transferir datos con ellos. La conexión se realiza con formato Uniform Resource Locator (URL). cURL sirve para realizar acciones sobre archivos que hay en Internet, soportando los protocolos de comunicación más comunes, como por ejemplo Hypertext Transfer Protocol (HTTP).

En lo que respecta a PHP, cURL no está integrado por defecto, de manera que para usar estas librerías se debe instalar el paquete libcurl.

Una consulta simple puede ser por ejemplo la siguiente Turland, M. (2010)(4) :

```
$c = curl_init('http://www.hidraulica.gob.ar/ema.
php?station=concordia');
curl_setopt($c, CURLOPT_RETURNTRANSFER, true);
$a = curl_exec($c);
curl_close($c);
```

En este caso, la variable $a contendrá el código fuente del sitio web indicado en curl_init. Para operar con HTML, PHP ofrece dos librerías que son GET y POST.

CRON en sistemas UNIX/LINUX

En los sistemas operativos basados en UNIX, como por ejemplo LINUX, existe un administrador regular de procesos en segundo plano llamado demonio que se ejecuta a intervalos regulares, por ejemplo, cada minuto, día, semana o mes. El nombre de este "demonio" es CRON. (Cron & Crontab, 2010) [5] y permitirá ejecutar de forma periódica el algoritmo de scraping.

Resultados

La primera parte del algoritmo consiste en extraer los datos de las diferentes fuentes u origenes. Los entornos de origen son usualmente bases de datos y/o ficheros, pero en ocasiones también pueden ser colas de mensajes, así como ficheros u otras fuentes estructuradas, semiestructuradas o

no estructuradas. (Castillo Montalvan, L. F. R. 2011) [6]. Los sistemas web de las estaciones automáticas E.M.A. son páginas desarrolladas en lenguaje HTML o superior, JavaScript o Extensible Markup Language (XML) que son lenguajes estructurados.

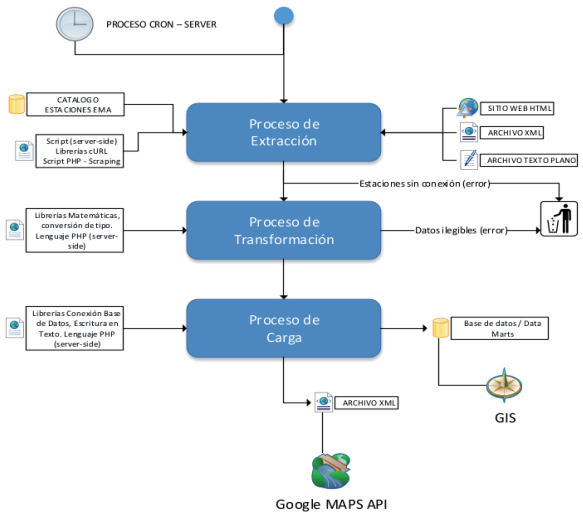

Figura 2: Diagrama de flujo del algoritmo completo.

Se detallan las implicancias en cada ítem con el correspondiente diagrama de flujo para el ítem de extracción, transformación y carga correspondiente a la figura 2.

Extracción

Paso 1. Obtener dirección web del inventario de estaciones EMA. El algoritmo intenta obtener de una tabla cargada en motor de base de datos, la primera dirección web o URL a procesar. En caso de obtener un error se descarta esa estación y se continúa con la siguiente de la lista. Algunos de los errores típicos son:

- Dirección web inexistente.
- Falla en la conexión a internet.
- Vencimiento del temporizador del protocolo de Internet.

El motor de base de datos para las pruebas se monta sobre sistema operativo LINUX y servidor web APACHE.

Paso 2. Identificar el lenguaje de programación.

Cada lenguaje maneja una estructura y palabras reservadas distintas, por lo tanto, se debe poder identificar qué tipo de origen es para poder extraer el dato. Cada tipo de lenguaje se identifica en su cabecera, por ejemplo HTML (<!DOCTYPE html>), JavaScript (<script type=”text/javascript”>) y XML (<?xml version=”1.0” encoding=”UTF-8” ?>). En caso de no poder identificar el lenguaje entonces se procede a buscar los valores en base a algún patrón conocido.

Paso 3. Identificar similitud en patrones.

La mayoría de los organismos y empresas poseen más de una estación meteorológica, esto significa que utilizan el mismo formato de presentación de datos o estilo de diseño en sus sitios web. En el relevamiento para realizar el inventario de estaciones meteorológicas se identifica las estaciones con estructura web similar para armar plantillas específicas que agrupen a varias E.M.A. El algoritmo procede a identificar si la estación pertenece a un grupo de estaciones conocidas, obtiene de una tabla de la base de datos la huella clave para identificar a la web procesada.

Paso 4. Obtener plantilla de búsqueda de palabras claves de la base de datos. Las palabras claves son necesarias para aplicar el scraping a las páginas web de las diferentes estaciones EMA. El algoritmo obtiene la cadena de caracteres de búsqueda de información de una tabla almacenada.

Paso 5. Extracción por scraping.

Utilizando la técnica de búsqueda de palabras claves o también denominado concordancia o coincidencias de expresiones regulares, se extraen datos significativos. Estos datos se analizarán posteriormente para validarlos.

Transformación

Para entender la necesidad de un proceso de transformación, se debe tener en cuenta que en un proceso de extracción, transformación y carga (ETL), se manejan fuentes diversas. Esta variedad de sitios web, en ocasiones de varios países, con diferentes idiomas y distintas unidades de medida, imposibilita o dificulta la posibilidad de realizar comparaciones si con anterioridad no se realizan conversiones y normalizaciones. De ahí la necesidad de los procesos de transformación. (PowerData, 2016) [7].

Paso 1. Eliminar datos fuera de fecha.

Se procede a eliminar todos aquellos datos leídos donde se identifique que la fecha de publicación no es la actual. Esto ocurre cuando el sitio web de la estación deja de publicar datos por la razón que fuere y permanecen en el sitio datos antiguos.

Paso 2. Eliminar datos con caracteres carentes de significado.

Se procede a eliminar todos aquellos datos leídos donde se identifique caracteres que no corresponden a la variable, por ejemplo un valor de temperatura que sea “#” significa que la estación no tiene ese dato procesado o que existe un falla en la misma.

Paso 3. Reformateo de datos.

Se procede a unificar los datos en unidades con decimal separado por punto. Es decir si se tiene un valor de temperatura que es 16,2 se formatea en 16.2.

Paso 4. Conversión de unidades.

Se procede a estandarizar los datos en grados Celsius, milímetros para las precipitaciones y hectopascal para la presión atmosférica. Muchos sitios contienen valores por ejemplo en grados Fahrenheit.

Carga

La última parte del proceso (ETL) es la fase de carga, el momento en el cual los datos procedentes de la fase de transformación son cargados en la base de datos.

Fundamentalmente, existen dos tipos de carga:

• Inserts: Es un sistema de acumulación simple consistente en el transporte de la información en grandes bloques de datos, previamente calculados en función, generalmente, de un valor sumatorio o de un promedio de la magnitud considerada. Se trata de la forma más sencilla y común de llevar a cabo un proceso de carga, pero tiene el inconveniente de que ante un accidente o problema, ya sea un corte de luz, un fallo del disco, etc., se pierde la consistencia de los datos, pudiéndose darse el caso de tener que repetir toda la carga.

• Loads: En este caso, la carga se realiza de forma más escalonada y segura. Para ello, el sistema agrupa la información de forma automática y transparente según distintas variables: ya sea por fechas, por un número determinado de registros, etc. Esta modalidad permite procesar el punto exacto hasta el que se ha realizado la carga, lo que supone que si se produce un fallo sólo hay que retomar el proceso de carga desde ese punto concreto, sin necesidad de repetirlo todo de nuevo.

Cuando se utiliza un sistema insert el nivel de consistencia se reduce, puesto que una falla puede obligar a una repetición íntegra del proceso. Sin embargo, la duración del tiempo de carga es menor.

En el caso particular de este trabajo, se utiliza un sistema loads, por las siguientes razones:

• Se asegura lo más posible la consistencia de los datos que se están cargando.

• Se prioriza la calidad de la carga antes que la rapidez.

• No se sobrecarga el servidor con operaciones asociadas a grandes cantidades de datos.

• Se atomiza los procesos por estación y por intervalo de tiempo.

Codificación del Algoritmo

El siguiente código documentado en lenguaje PHP, permite extraer datos de una estación que publica en formato XML. Se realiza además el proceso de transformación de unidades. La programación fue simplificada para ejemplificar la extracción de una estación.

```
<?
//OBTENER URL
//Devuelve un manipulador de cURL si todo fué bien, FALSE si hay errores.
$c =
curl_init('http://api.wunderground.com/weatherstation/WXCurrentObXML.asp?ID=IBUENOSA73');
//Establece una opción en la sesión del recurso cURL.
curl_setopt($c, CURLOPT_RETURNTRANSFER, true);
//Ejecuta la sesión cURL que se le pasa como parámetro.
$a = curl_exec($c);
//Esta función cierra una sesión CURL y libera todos sus recursos.
curl_close($c);
//Latitud
$ft = strpos($a, '<latitude>') + strlen('</latitude>');
$latitud = (float)($a[$ft-1].$a[$ft+0].$a[$ft+1].$a[$ft+2].$a[$ft+3].$a[$ft+4].$a[$ft+5].$a[$ft+6].$a[$ft+7].$a[$ft+8]);
//Longitud
$ft = strpos($a, '<longitude>') + strlen('</longitude>');
$longitud = (float)($a[$ft-1].$a[$ft+0].$a[$ft+1].$a[$ft+2].$a[$ft+3].$a[$ft+4].$a[$ft+5].$a[$ft+6].$a[$ft+7].$a[$ft+8]);
//Altura en Pies
$ft = strpos($a, '<elevation>') + strlen('</elevation>');
$altura_ft = (float)($a[$ft-1].$a[$ft+0].$a[$ft+1]);
//Transformación Pies a Metros
$altura_m = ($altura_ft/3.2808);
//Última actualización
$ft = strpos($a, '<observation_time_rfc822>') + strlen('</observation_time_rfc822>');
$fecha = ($a[$ft-1].$a[$ft+0].$a[$ft+1].$a[$ft+2].$a[$ft+3].$a[$ft+4].$a[$ft+5].$a[$ft+6].$a[$ft+7].$a[$ft+8].$a[$ft+9
]);
//Temp C
$ft = strpos($a, '<temp_c>') + strlen('</temp_c>');
$temp = (float)($a[$ft-1].$a[$ft+0].$a[$ft+1].$a[$ft+2]);
//Precipitación en in (pulgadas)
$ft = strpos($a, '<precip_today_in>') + strlen('</precip_today_in>');
$lluvia_in = (float)($a[$ft-1].$a[$ft+0].$a[$ft+1].$a[$ft+2]);
//Transformación In a MM
$lluvia_mm = ($lluvia_in/0.039370);
//Intensidad
$ft = strpos($a, '<precip_1hr_in>') + strlen('</precip_1hr_in>');
$intensidad_in = (float)($a[$ft-1].$a[$ft+0].$a[$ft+1].$a[$ft+2]);
//Transformación In a MM
$intensidad_mm = ($intensidad_in/0.039370);
//Presion en MB
$ft = strpos($a, '<pressure_mb>') + strlen('</pressure_mb>');
$presion_mb = (float)($a[$ft-1].$a[$ft+0].$a[$ft+1].$a[$ft+2].$a[$ft+3].$a[$ft+4]);
//Transformación MB a Hpa
$presion_hpa = ($presion_mb);
//Humedad Relativa
$ft = strpos($a, '<relative_humidity>') + strlen('</relative_humidity>');
```

```
$humedad = (float)($a[$ft-1].$a[$ft+0].$a[$ft+1].$a[$ft+2]);
//Resultado
echo "LATITUD : ".$latitud."<br>";
echo "LONGITUD : ".$longitud."<br>";
echo "ALTURA MSNM : ".$altura_m."<br>";
echo "FECHA : ".$fecha."<br>";
echo "TEMPERATURA C° : ".$temp."<br>";
echo "PRECIPITACIÓN DIA MM : ".$lluvia_mm."<br>";
echo "PRECIPITACIÓN 1 HORAMM : ".$intensidad_mm."<br>";
echo "PRESIÓN ATM HPA : ".$presion_hpa."<br>";
echo "HUMEDAD RELATIVA : ".$humedad."<br>";
?>
```

Conclusiones

La necesidad de obtener datos meteorológicos de distintas estaciones automáticas, la falta de un lugar único o repositorio unificado desde donde poder consultarlos, sumado a la escasez de los mismos, motivó el trabajo de investigación. (Héctor R. López, 2017) [8].

Se investigaron distintas alternativas para poder extraer datos publicados por estaciones meteorológicas en tiempo real, optándose por herramientas de extracción web en Scraping.

A través de herramientas scraping, se logró enfocar el objetivo, desarrollar una técnica de extracción, transformación y carga, también conocida como ETL. Mediante éste instrumento se consultaron páginas HTML de estaciones E.M.A. con diversos formatos, distintos orígenes y se extrajeron variables como temperatura, precipitaciones, humedad, entre otras.

Sistemas operativos actuales, como LINUX, permiten programar las lecturas de datos en intervalos preestablecidos, automatizando la ejecución del algoritmo desarrollado.

Se construyó un inventario de estaciones reportando en todo el territorio Argentino. El inventario, de más de 500 estaciones, abarca organismos oficiales, empresas, cooperativas y terceros.

El lenguaje de programación PHP fue una herramienta clave para el desarrollo, ya que a través de sus librerías cURL se codificó el núcleo del algoritmo.

cURL, permitió abrir conexiones en una amplia variedad de protocolos y conectar a distintos entornos, como el web o XML.

Se normalizaron y almacenaron esas variables extraídas en un motor de base de datos para su posterior consulta y además para que puedan ser utilizadas por sistemas geográficos.

Se desarrolló un sistema de monitoreo geográfico con la herramienta API Google Maps, disponible para expertos, analistas y científicos.

Agradecimientos

A la Universidad Nacional de San Juan, la Facultad de Ciencias Exactas, Físicas y Naturales, al Departamento de Posgrado y al equipo de docentes de la Maestría en Informática por compartirme sus conocimientos y experiencia.

Al Magister Ing. Raúl Klenzi que, como director de esta tesis, me ha orientado, apoyado y corregido en mi labor académica con un interés y una entrega incansable.

Al Doctor Ing. Oscar Dölling una persona que sabe inspirar la pasión por la investigación y el trabajo.

A mi esposa, mi hijo, mis padres, mi hermana, que siempre me apoyaron incondicionalmente y son los pilares de mi vida.

Bibliografía

1. **Turland, M. PHP. Marco Tabini & Associates.** Inc. 2010.
2. **Penman, R. B., Baldwin, T., & Martinez, D.** *Web scraping made simple with sitescraper.* 2009.
3. **Mehlführer, A.** *Web scraping: a tool evaluation.* 2009.
4. **Cron & crontab.** *Visible Body: Cron & Crontab. 2010.* Recuperado de http://blog.desdelinux.net/cron-crontab-explicados/ - Accedido 2016.
5. **Castillo Montalvan, L. F. R.** *Aplicación de la metodología de Amón y Jiménez para asegurar la calidad de los datos en la construcción del ETL durante la implementación de un datamart para la Empresa MC Express de la ciudad de Chiclayo.* 2011.
6. **PowerData.** *Visible body: Procesos ETL: Transformación. ¿En qué Consiste?*. Recuperado de: http://blog.power-data.es/el-valor-de-la-gestion-de-datos/bid/312589/Procesos-ETL-Transformaci-n-En-qu-Consiste - . 2016. Accedido 2016.
7. **Lépez, Héctor R.** *Técnica de Extracción, Transformación Y Carga de datos de Estaciones Meteorológicas.* Maestría En Informática. 2017.

Recibido: 07/08/2018.
Aprobado: 03/10/2018.