Ejes temáticos: Visibilidad Académica
Tipo de trabajo: ponencia

Implementación de algoritmo para la extracción de datos estructurados de perfiles en Google Académico

Danny Murillo [1], Dalys Saavedra [2], Dalys Huriviades Calderón [3]
[1,2,3] Universidad Tecnológica de Panamá, Panamá
[1] danny.murillo@utp.ac.pa, [2] dalys.saavedra@utp.ac.pa,
[3]huriviades.calderon@utp.ac.pa

Danny Murillo, licenciado en Ingeniería de Sistemas Computacionales, Maestría en Gerencia informática con énfasis en seguridad computacional, labora desde hace 15 años en la Universidad Tecnológica de Panamá, es investigador en las áreas de ingeniería web, minería de datos, indicadores de ranking universitario. Ha sido profesor instructor en los temas de diseño web, programación web, minería de datos, lenguaje R, análisis predictivo de datos, redes científicas, repositorios institucionales. Es coordinador del proyecto UTP-Ridda2, repositorio institucional de acceso abierto de la UTP y del proyecto de visibilidad y alcance de la producción científica y académica de la UTP.

Dalys Saavedra, Ingeniería en Sistemas Computacionales, Maestría en énfasis en seguridad de la Universidad de Zaragoza, España, labora desde hace 12 años en la Universidad Tecnológica de Panamá como administradora de recursos informáticos y administradora de proyectos nacionales e internacionales en el área de tecnología. Ha sido instructora en cursos de Excel, uso de Open Journal System, roles del portal de revistas y perfiles en Google Scholar. Es administradora del Portal de revistas digitales de la UTP y participa en el proyecto de visibilidad y alcance de la producción científica y académica de la UTP.

Huriviades Calderón, Ingeniero en Sistemas Computacionales, labora como investigador en la Dirección de Investigación de la Universidad Tecnológica de Panamá desde hace 2 años. Es miembro del grupo de investigación GISES en salud electrónica y servicios en la nube. Ha participado en proyectos internacionales, es editor de la revista I+D Tecnológica. Es administrador del Repositorio Institucional de la UTP encargado de implementación y políticas de interoperabilidad y preservación, participa en el proyecto de visibilidad y alcance de la producción científica y académica de la UTP.

RESUMEN

Este trabajo muestra el desarrollo e implementación de un algoritmo para extraer datos de perfiles y publicaciones de Google Académico (GA) utilizando Web Scraping, técnica no estructurada de minería de datos que escanea los datos de una página web. El código del algoritmo se crea utilizando el lenguaje R el cual nos permite personalizar la extracción de datos implementando funciones de extracción de perfiles y publicaciones de una universidad. En las pruebas de extracción de

datos realizadas con las herramientas web y online se logró un promedio de 2 a 8 horas para extraer un promedio de 55 perfiles y 1400 publicaciones, mientras que con el algoritmo se logra extraer la misma cantidad de perfiles y publicaciones en 4 minutos con datos estructurados en formato de tabla que pueden ser exportadas para su posterior uso. Estas pruebas fueron realizadas en un periodo de 1 año, depurando errores y mejorando tanto el tiempo de extracción de los datos de salida. Una de las limitantes del algoritmo es que en universidades con más de 2,000 perfiles, este, es bloqueado por GA debido a que el tiempo de extracción aumenta y considera que es un robot o araña quien escanea los datos, es por ello que se trabaja en mejorar el proceso de extracción. El trabajo realizado permite que este algoritmo sea una herramienta para quienes realizan análisis de datos de indicadores científicos o para quienes realizan análisis bibliométricos de revistas académicas y científicas con perfiles en GA.

Palabras claves: Google Académico, escaneado web, Minería web, Lenguaje R, análisis de datos.

Keywords: Google Scholar, Web Scraping, Web Mining, R Lenguage, data analysis.

INTRODUCCIÓN

La sobreabundancia de información en Internet, es uno de los principales componentes de su éxito, sin embargo el tratamiento de esta exige una enorme cantidad de tiempo y energía a fin de cribar la calidad de los datos sumergidos en tan enorme repositorio (Sánchez Carballido, 2011). Estos datos están organizados, estructurados y visibles en páginas web, pero, no siempre es posible poder extraerlos con la rapidez o la estructura deseada para su posterior análisis.

Uno de estos ejemplos es la web de Google Académico (GA), un buscador de Google lanzado en noviembre en 2004 enfocado en visibilizar la producción científica académica de los docentes e investigadores enfocados en indicadores bibliométricos como el hindex (Dávalos-Sotelo, 2015) citaciones, número de publicaciones y fechas de indexación siendo actualmente objeto de análisis científico (Mugnaini & Strehl, 2008).

Aunque resulta de interés los datos generados por GA, su problema radica en que no es posible extraer los datos de forma simple, ya que no cuenta con alguna opción para descargar los datos de perfiles o publicaciones siendo aún más complejo cuando se trata de varios perfiles de investigadores de una universidad.

La necesidad de extraer estos datos de GA se debe a que diversos estudios han surgido buscando analizar la potencialidad, confiabilidad y cobertura de esta plataforma como instrumento de recuperación de información, siendo posible localizar citas emitidas por documentos no cubiertos por otras bases de datos como (Giustini & Kamel Boulos, 2013) preprints, tesis, informes, libros, repositorios o publicaciones de revistas en indexadores regiosnales (Orduña-Malea, 2016), el uso de estos datos cambia el modelo adoptado por bases de datos tradicionales y de pago como Scopus, Scimago (SJR), Web of Science (WoS), ampliando el acceso

del universo de documentos indexados, lo que permite una mejor evaluación de la ciencia (Ortega, 2015).

Por otro lado se hace necesario que tanto investigadores como las universidades puedan tener acceso a información de indicadores científicos – académicos que permita analizar y evaluar el impacto de sus publicaciones en esta plataforma.

El objetivo de este trabajo muestra el desarrollo e implementación de un algoritmo para extraer datos de Google Académico utilizando la técnica de Web Scraping, técnica no estructurada de minería de datos que escanea los datos de una página web (Directions, Mining, Further, & Education, 2012) y permite estructurarla en formato de tabla y por campos, la cual puede ser luego exportada o manipulada para hacer análisis de datos (Shi, Liu, Shen, Yuan, & Huang, 2015) (Bharanipriya & Prasad, 2011).

El algoritmo se realizó utilizando el lenguaje de programación de R de código abierto, lenguaje de script que no requiere ser compilado para ser ejecutado y tiene similitud con otros lenguajes orientado a objetos(Cotton, 2013).

MATERIALES Y METODOLOGÍA

Para realizar las pruebas de extracción de datos utilizando web scraping se utilizaron los siguientes métodos:

Copiar y Pegar: no es un método de Web Scraping, pero es la forma más común de extraer datos de un sitio web, el proceso consistió en copiar y pegar cada dato del perfil y las publicaciones en una tabla de Excel, seleccionando solo el dato que se necesitaban.

Web scraping Local: se instaló extensión SCRAPER de Google Chrome para seleccionar el bloque de datos de una página web, extrae los datos que tengan el mismo patrón de la clase HTML seleccionada, solo permite extraer los datos una página web por vez el ciclo de repetición de Web Scraping lo debe hacer el usuario.

Web scraping Local Software: se instaló el software FMiner, que permitió abrir la página web en la aplicación y seleccionar los nodos html, creando un diagrama de flujo de datos de la página web, el proceso es semi-automático, ya que el usuario debe escoger los datos que desea guardar (Fminer, 2015).

Web scraping Online: se utilizó Import.io una aplicación web que analiza automáticamente la estructura de la página web y muestra los datos en formato de tabla. (Import.io, 2016).

Algoritmo en R v1: se seleccionó como base del algoritmo el paquete scholar el cual extrae datos de GA por perfil, creado por James Keirsted en el 2015, actualizado en junio de 2016.

Algoritmo en R v2: se creó el algoritmo desde cero utilizando el lenguaje de programación R y la interface de R studio para Windows, se utilizaron los paquetes, rvest, stringr, tidy, xml2, plyr, wordcloud, dplyr, plot.

Algoritmo en R v3: se usó la base del algoritmo v2, programación vectorizada, corrección de extracción de datos de perfiles con muchas publicaciones y creación de automatización de funciones.

Algoritmo en R v4: base del algoritmo v3, adaptación de nuevas etiquetas de GA, corrección de errores de perfiles y publicaciones en blanco, corrección de paginación en los perfiles, creación de funciones de agrupación citas y publicaciones.

Hardware: computadora con Windows 10 de 64 Bits, memoria RAM de 12 GB. La velocidad de Internet en periodo de pruebas fue de 5 a 35 Mb de descarga

Población

Para la primera prueba de extracción y la comparación de velocidad de extracción de datos utilizando diferentes métodos de extracción y algoritmo en R v1 que utiliza el paquete scholar, se seleccionaron 5 perfiles de universidades con perfil en GA con un promedio de 55 perfiles afiliados: Universidad Francisco Marroquin (UFM), Guatemala, Escuela Superior Politécnica del Litoral (ESPOL), Ecuador, Universidade Regional de Blumenau (FURB), Brasil, Universidad Tecnológica de Panamá (UTP), Panamá, Universidad de La Habana (UH), Cuba.

Para la segunda prueba de extracción de perfiles se utiliza el Algoritmo en R v1 con el paquete scholar y el algoritmo v2 totalmente programado en R, se seleccionaron 10 nuevas Universidades en GA con un promedio de 800 perfiles afiliados, Universidad de la República (UDELAR), Universidad de Costa Rica (UCR), Université de Franche-Comté (UFC), Universidad de Antioquia (UDEA), Universidad de Chile (UCHILE), Universidad Nacional Autónoma de México (UNAM), Universidad de Osaka (OSAKAU), University of Edinburgh (UED), Universidad Politécnica de Valencia (UPV), University of Illinois at Urbana-Champaign (UILLINOIS).

Para la prueba de extracción utilizando el algoritmo v4, actualización de etiquetas de Google Académico se seleccionaron del sitio web de ranking de universidades, Webometrics, encontrando que existían 362 universidades de Centro América y El Caribe de 18 países con dominios URL a su página web. Para el estudio solo se consideraron solamente las universidades con perfil y afiliación en GA.

Procedimiento para creación de algoritmo en R

Para la creación del algoritmo en R se siguió el procedimiento mostrado en la figura1, se cargó las librerías y funciones de R a utilizar. El proceso de extracción requiere conocer la url de afiliación de la universidad. Para la extracción de los datos se el código fuente de la página web un etiquetado ordenando de nodos que se conoce normalmente como DOM (Document Object Model). Cada nodo representa una etiqueta HTML y juntas puede representarse como ramas ordenadas y rotuladas. La jerarquía de árbol representa los diferentes niveles de anidamiento de los elementos que constituyen la página Web, la idea detrás del modelo es que las páginas web que contiene etiquetas HTML, se muestren como texto y palabras claves que puede ser interpretado por el navegador (Ferrara, De Meo, Fiumara, & Baumgartner, 2014).

Se analizó la estructura del DOM de los bloques de los perfiles de la afiliación de la universidad en GA, identificando patrones repetitivos en los datos como el nombre, citaciones, palabras claves. Se separaron cada uno de los nodos HTML que contenían los datos de los elementos individuales y se almacenaron en variables para luego agruparlas y almacenarlos en una tabla en R llamada data.frame, esta permite almacenar diferentes tipos de datos

Los datos en formato de tabla de las url de perfiles se utilizan para extraer los detalles del perfil como palabras, claves, numero de citaciones y hindex (Murillo & Saavedra, 2017). pero las url son también entrada de datos para poder iniciar el proceso de escáner página de publicaciones por perfil.

La extracción de publicaciones se repite por cada perfil y se van almacenando en una tabla de todas las publicaciones por perfil de dicha universidad. Luego los datos pueden ser usados para crear gráficas o agrupas número de citaciones o publicaciones por perfil.

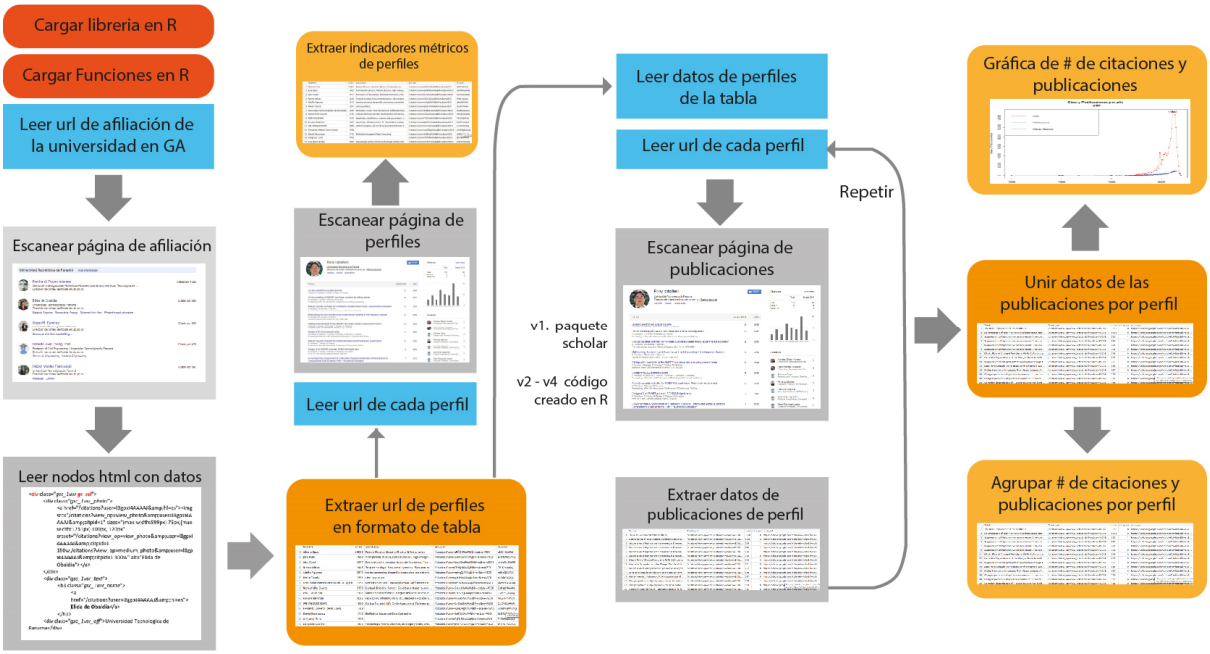


Figura 1. Esquema del funcionamiento del algoritmo en R

Prueba de velocidad de extracción: para probar la velocidad de extracción del algoritmo vs los otros métodos de extracción se utilizaron los datos de las 5 Universidades seleccionadas con 55 perfiles y 1400 publicaciones como promedio, luego se contabilizó el tiempo en minutos de cada método.

Prueba de algoritmos v1 y v2: con el objetivo de realizar una prueba entre el algoritmo v1 utilizando el paquete scholar y el algoritmo creado totalmente en R, se utilizaron para las pruebas 10 universidades con promedio de 880 perfiles.

Actualización de algoritmo v4: este algoritmo se actualizo en enero de 2018, debido a que Google Académico cambio su interface y las etiquetas html y clases en CSS que contenían los datos de perfiles y publicaciones, además la forma en que se mostraban los detalles de las publicaciones fue cambiada de mostrar los datos en una página aparte a mostrar los datos en una pantalla modal, lo que genero realizar esta actualización, la cual también permitió generar otras funciones como hacer análisis individual de un perfil y agrupar datos de publicaciones y citaciones por perfil, como también extraer los coautores y la publicación.

RESULTADOS

Para la prueba de extracción del algoritmo se utilizó la v1 el cual utilizaba el paquete scholar, para el proceso se contabilizó de forma manual el número de perfiles y publicaciones de cada universidad antes de su extracción. En la tabla 1 se muestra el número de perfiles y publicaciones extraídas de las universidades el cual es igual al número de publicaciones contabilizadas manualmente, mostrando un tiempo promedio de extracción por universidad de 3.8 minutos.

Tabla 1

Tiempo en minutos de extracción de perfiles y publicaciones de algoritmo en R

Universidad	País	Perfiles		Publicaciones		Tiempo
		manual	extraídos	manual	extraídas	
UFM	Guatemala	14	**14**	393	**393**	3
ESPOL	Ecuador	67	**67**	1061	**1061**	3
FURB	Brasil	38	**38**	1360	**1360**	4
UTP	Panamá	77	**77**	1434	**1434**	4
UH	Cuba	79	**79**	2758	**2758**	5

En la Figura 2 se muestra los perfiles vistos en GA y como se muestra la extracción de los perfiles en formato de tabla utilizando el algoritmo, esta tabla se puede exportar a formato .csv o a formato .xls para su posterior uso o análisis.

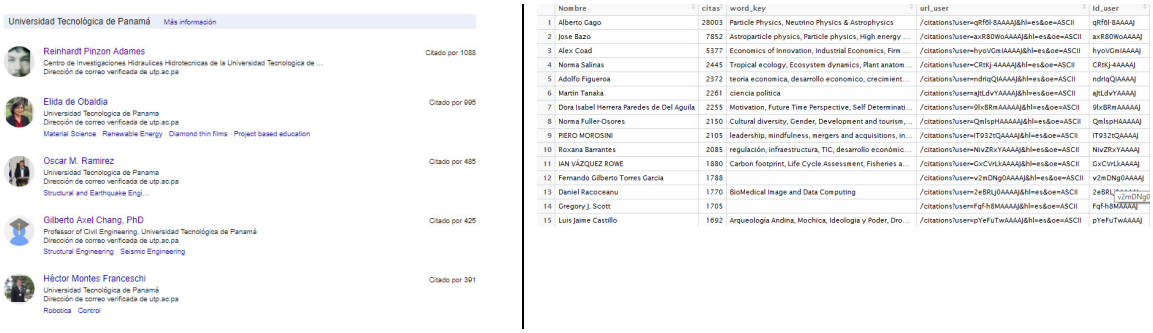
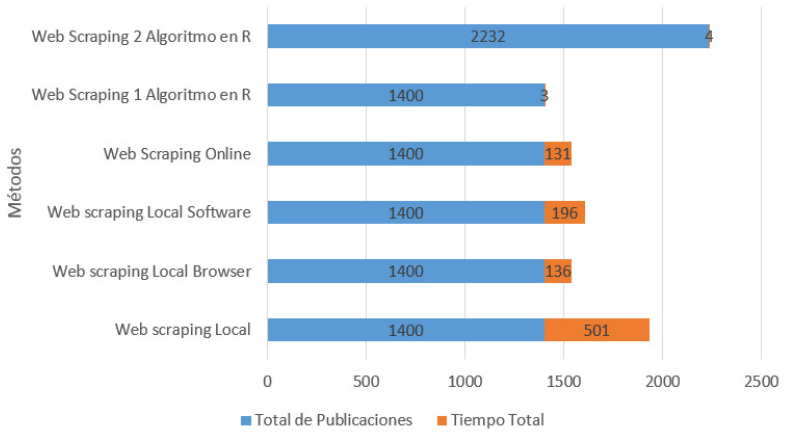


Figura 2. Perfil en Google Académico (izquierda)
Exracción de perfiles en formato de tabla (derecha)

El resultado de velocidad de extracción de los métodos vs el algoritmo se muestra en la gráfica 1, el algoritmo en R, método “1 algoritmo en R” es de 3 minutos incluyendo perfiles y publicaciones. El método “2 algoritmo en R”, es el mismo algoritmo, pero se incluyó la Universidad de la República del Uruguay (UDELAR) con 182 perfiles y 6388 publicaciones. El tiempo de este método fue de 4 minutos, también inferior al tiempo de los otros métodos, aún mejor es la eficiencia del algoritmo considerando que los datos extraídos del algoritmo están estructurados en formato de tabla, lo que no es posible con los otros métodos.



Gráfica 1. Comparación Tiempo en minutos de los Método de Scrapear de Perfiles y Publicaciones en GA

En la tabla 2 muestra el resultado de la prueba de extracción con las 10 universidades con promedio de 800 perfiles, aunque el algoritmo v2 programado totalmente en R muestra un tiempo menor de 38 minutos extrayendo los perfiles de las 10 universidades versus los 122 minutos del algoritmo v1, realmente el resultado de mayor impacto es que el número de perfiles extraídos de universidades entre los dos algoritmo es diferente, siendo el algoritmo v2 el que extrae la cantidad correcta de perfiles, según se contabilizaron manualmente antes de la prueba .

Tabla 2

Perfiles extraídos por universidad en GA utilizando algoritmo con paquete "scholar" y algoritmo sin paquete scholar

		Con paquete Scholar		Sin paquete Scholar	
Universidad	**#Perfiles correctos**	**#Perfiles**	**Tiempo (minutos)**	**#Perfiles**	**Tiempo (minutos)**
UDELAR	182	182	2	182	1
UCR	230	230	2	230	1
UFC	119	119	2	119	1
UDEA	383	383	3	383	1
UCHILE	**566**	**137**	5	**566**	2
UNAM	**1329**	**138**	11	**1329**	5
OSAKAU	**460**	**140**	4	**460**	1
UED	**1471**	**147**	14	**1471**	5
UPV	**794**	**387**	11	**794**	3
UILLINOIS	**2555**	**2250**	62	**2555**	12
	8364	**4388**	**122**	**8364**	**38**

El algoritmo con el paquete Scholar mostró errores en el número de extracción de perfiles y publicaciones. En las primeras 4 Universidades mostradas en la tabla 2 se verificó de forma manual que los resultados de cantidad de perfiles y publicaciones es el indicado. En las otras universidades los perfiles que tenían más de 100 publicaciones se extrajo 0 publicaciones, al verificar los perfiles algunos casos tenían hasta 2000 publicaciones. En la extracción de datos de la UNAM de 566 perfiles solo se extrajeron 137 perfiles con el número de publicaciones correctas, en OSAKAU de los 1329 solo 140. El problema que encontramos es que la función get_publications del paquete scholar contiene una variable (FLUSH=false) que extrae datos en caché, cuando se habilitó a (FLUSH=true), algunos perfiles que mostraron 100, cambiaron su valor a 1000 a la hora de volver hacer la extracción, algunos perfiles que habían sido extraídos de forma correcta, pasaron a tener 0 publicaciones, por lo que el valor de (FLUSH) no permite scrapear los datos cuyos perfiles tengan más de 100 publicaciones.

Las pruebas de extracción de datos de 30 universidades de Centro américa y el caribe con perfil en Google Académico utilizando la v4 del algoritmo se muestran en la tabla 3, donde las actualizaciones realizadas permitieron realizar la extracción total de perfiles y publicaciones, sin embargo en algunas ocasiones el proceso fue bloqueado por Google académico por un periodo de 1 hora, los errores mostrados fueron, Error 403 Forbidden, Error 503 Service Unavailable, 301 Moved Temporarily, 401 Unauthorized, 404 Not Found, 408 Request Timeout, 429 Too Many Requests. Por otro lado este nuevo algoritmo y la nueva estructura de GA han permito una reducción de 5.56 minutos de la v3 a 5.13 del algoritmo v4.

Tabla 3

Perfiles extraídos de Universidades en Centro américa y el caribe utilizando algoritmo v4 en R

País	Universidad	perfiles	publicaciones	Tiempo (minutos)			
				perfiles	Publica	v4	v3
Puerto Rico	Universidad de Puerto Rico	296	13541	3	12	15	17
Costa Rica	Universidad de Costa Rica	245	11110	3	10	13	14
Jamaica	University of the West Indies	337	9896	5	8	13	14
Trinidad and Tobago	University of the West Indies at St Augustine	205	7877	3	6	9	10
Costa Rica	Tecnológico de Costa Rica	264	6918	3	6	9	10
Costa Rica	Universidad Nacional de Costa Rica	115	5324	3	6	9	10
Costa Rica	Centro Agronómico Tropical de Investigación y Enseñanza	56	3846	1	4	5	7
Cuba	Universidad de la Habana	119	3453	3	4	7	8
Cuba	Instituto Superior Politécnico José Antonio Echeverría	87	1888	2	3	5	6
Cuba	Universidad de Ciencias Informáticas	74	1841	2	3	5	5
Cuba	Universidad Central Marta Abreu de la Villas	61	1793	1	3	4	4
Cuba	Universidad de Camaguey	47	1636	1	3	4	4
Panamá	Universidad Tecnológica de Panamá	87	1592	2	3	5	5
Puerto Rico	Ponce School of Medicine	12	1273	1	3	4	4
Honduras	Escuela Agrícola Panamericana Zamorano	15	1273	1	3	4	4
Cuba	Instituto Nacional de Ciencias Agrícolas	33	1171	1	3	4	4
Guatemala	Universidad del Valle de Guatemala	24	753	1	2	3	4
Puerto Rico	Universidad Central del Caribe	8	642	1	2	3	3
Costa Rica	Universidad Estatal a Distancia Costa Rica	26	600	1	2	3	3
Cuba	Universidad de Oriente Santiago de Cuba	37	580	1	2	3	3
Cuba	Instituto Superior de Tecnologias y Ciencias Aplicadas	18	535	1	2	3	3
Cuba	Universidad de Pinar del Rio	12	517	1	2	3	3
República Dominicana	Pontificia Universidad Católica Madre y Maestra	42	480	1	2	3	3
Guatemala	Universidad de San Carlos de Guatemala	27	441	1	2	3	3
Guatemala	Universidad Francisco Marroquín	13	397	1	2	3	3
Costa Rica	INCAE Business School	9	392	1	2	3	3
El Salvador	Universidad Don Bosco El Salvador	20	202	1	2	3	3
Jamaica	University of Technology Jamaica	12	178	1	2	2	3
Puerto Rico	Universidad Politécnica de Puerto Rico	8	159	1	2	2	2
Costa Rica	Universidad Centroamericana José Simeón Cañas	7	130	1	2	2	2
		2316	**80438**				

CONCLUSIÓN

Los resultados obtenidos de las pruebas de métodos de Web Scraping muestran que esta técnica es una alternativa funcional para extraer datos de un sitio web pasando de tener datos no estructurados a datos estructurados.

La creación de las diferentes versiones del algoritmo en R ha tomado un periodo de un año entre su desarrollo, pruebas e implementación, sin embargo a resultado ser la mejor opción en las pruebas de velocidad de extracción realizadas a las casi 45 universidades con afiliación, durante el proceso de extraer datos de perfiles y publicaciones el cual permite reutilizar estos datos para su análisis.

Sin bien existe un beneficio en el uso del algoritmo en ser mejor en la extracción de los datos inclusive comparado con otros paquetes desarrollados como scholar, tiene la limitante que es bloqueado por Google Académico si es usado en largos periodos de tiempo al ser considerado un robot, por lo que es necesario esperar un periodo de tiempo para su posterior uso.

La realización de este algoritmo permite tener una herramienta que puede ser utilizada por otras universidades con perfiles en Google Scholar o por investigadores para poder llevar métricas o indicadores científicos de sus investigadores.

Uno de los objetivos a largo plazo es hacer análisis de los perfiles de las revistas científicas con perfil en Google Académico.

Enlace del código del Algoritmo desarrollado en el Lenguaje R:
https://bitbucket.org/dannymu/ejemplos-de-r

BIBLIOGRAFÍA

Bharanipriya, V., & Prasad, V. K. (2011). Web Content Mining Tools : a Comparative Study. *International Journal of Information Technology and Knowledge Management*, *4*(1), 211–215.

Borrego, F. (2017). Alternativas para realizar web scraping. Retrieved from http://felicianoborrego.com/alternativas-para-realizar-web-scraping/

Cichini, K. (2012). GScholarScraper_3.1. Retrieved from https://github.com/gimoya/theBioBucket-Archives/blob/master/R/Functions/GScholarScraper_3.1.R

Cotton, R. (2013). *Learning R, O´RELLY*.

Danny Murillo, D. S. (2016). Implementación de Plataforma Digital de Revistas Académicas y Científicas electrónicas en la Universidad Tecnológica de Panamá para mejorar su visibilidad a nivel nacional e internacional. In *Tecnología, innovación e investigación en los procesos de enseñanza-aprendizaje* (pp. 936–947). Retrieved from https://dialnet.unirioja.es/servlet/articulo?codigo=6035234

Dávalos-Sotelo, R. (2015). Una forma de evaluar el impacto de la investigación

científica, *21*, 7–16.
Directions, D. I., Mining, T., Further, U. K., & Education, H. (2012). The Value and Benefits of Text Mining, (March).
Extension Google Chrome. (2015). Scraper. Retrieved from https://chrome.google.com/webstore/detail/scraper/mbigbapnjcgaffohmbkdlecaccepngjd?utm_source=chrome-app-launcher-info-dialog
Ferrara, E., De Meo, P., Fiumara, G., & Baumgartner, R. (2014). Web data extraction, applications and techniques: A survey. *Knowledge-Based Systems*, *70*, 301–323. https://doi.org/10.1016/j.knosys.2014.07.007
Fminer. (2015). FMiner Scraping. Retrieved from http://www.fminer.com/
Giustini, D., & Kamel Boulos, M. N. (2013). Google Scholar is not enough to be used alone for systematic reviews. *Online Journal of Public Health Informatics*, *5*(2), 0–4. https://doi.org/10.5210/ojphi.v5i2.4623
Import.io. (2016). Import.io. Retrieved from https://www.import.io/
Keirstead, J. (2015). Package Scholar. Retrieved from https://cran.r-project.org/web/packages/scholar/index.html
Mugnaini, R., & Strehl, L. (2008). Recuperação e impacto da produção científica na era google: uma análise comparativa entre o google acadêmico e a web of science 10.5007/1518-2924.2008v13nesp1p92. *Encontros Bibli: Revista Eletrônica de Biblioteconomia e Ciência Da Informação*, *13*(1), 92–105. https://doi.org/10.5007/1518-2924.2008v13nesp1p92
Murillo, D., & Saavedra, D. (2017). Web Scraping de los Perfiles y Publicaciones de una Afiliación en Google Scholar utilizando Aplicaciones Web e implementando un Algoritmo en R. In *Congreso AMITIC* (pp. 1–8).
Orduña-Malea, E. (2016). Aplicaciones métricas de Google Scholar para la evaluación del impacto científico, (April), 29–30. Retrieved from https://www.researchgate.net/publication/299537816_Aplicaciones_metricas_de_Google_Scholar_para_la_evaluacion_del_impacto_cientifico?pli=1&loginT=gzGd1_JXCLgdk0SLIgG-ZcxvyojrcmiGxZsZW3jbvBg&uid=TGfrsN61Z891PFiNYhQuWopMEZKDt33BjYqC&cp=re375_fw_sl2_p1001&ch
Ortega, J. L. (2015). Diferencias y evolución del impacto académico en los perfiles de google scholar citations: una aplicación de árboles de decisión. Retrieved from http://redc.revistas.csic.es/index.php/redc/article/view/905/1283
Sánchez Carballido, J. R. (2011). Perspectivas de la información en Internet: ciberdemocracia, redes sociales y web semántica. *Zer-Revista de Estudios de Comunicación*, *13; n.° 25*, 61–81. Retrieved from http://www.ehu.eus/ojs/index.php/Zer/article/view/3574
Shi, S., Liu, C., Shen, Y., Yuan, C., & Huang, Y. (2015). AutoRM: An effective approach for automatic Web data record mining. *Knowledge-Based Systems*, *89*, 314–331. https://doi.org/10.1016/j.knosys.2015.07.012

Requerimientos de equipo técnico para la presentación del trabajo: computadora, proyector, software R Studio, conexión a internet.