Implementación de algoritmo en el Lenguaje R para extraer los datos de los Perfiles en Google Scholar utilizando la técnica web Scraping de Minería de datos

Danny Murillo, Dalys Saavedra
Vicerrectoría de Investigación Postgrado y Extensión, Universidad Tecnológica de Panamá
Panamá, Panamá
danny.murillo@utp.ac.pa
dalys.saavedra@utp.ac.pa

Abstract— Este articulo muestra diferentes pruebas realizadas para extraer datos de los perfiles y publicaciones de una afiliación en Google Scholar utilizando la técnica de Web Scrpaing de minería de texto no estructurada. El objetivo es medir la facilidad de extracción de estos datos con esta técnica, llegando a la implementación de un algoritmo en el lenguaje R para automatizar el proceso, estructurar los datos y disminuir el tiempo de scraping. Estas pruebas se hicieron a 15 Universidades con diferente cantidad de perfiles y publicaciones. La realización de este algoritmo permitirá la extracción de datos a cualquier afiliación, aunque todavía hay elementos que se pueden mejorar para que el algoritmo sea óptimo, pero hemos de concluir que según las pruebas realizadas el método de web scripting es funcional para poder extraer datos de un sitio web.

I. INTRODUCCIÓN

Visibilizar la producción científica y académica de una Universidad es una buena práctica para la construcción de una identidad académica online si queremos que la información que generamos a través de artículos de investigación puedan tener un alcance Nacional e Internacional en la web [1]. Según Tim Berner-Lee, la idea de la web era diseñar un espacio de trabajo colaborativo que facilitará el flujo de información [2], con el paso del tiempo, la vinculación a la web de páginas html, enlaces de hipertexto, la WWW dejó de ser una red de enlaces entre páginas y documentos evolucionando a una red de datos [3]. Estos documentos y páginas son datos que genera cada Universidad que a su vez generan conocimiento, sin embargo, si como Universidad no medimos el impacto que tienen estos trabajos en la web, no sabremos si estamos logrando el alcance que queremos.

Una de las formas más comunes de visibilizar la información de investigación y académica es utilizando portales de Revistas y repositorios institucionales, estas Plataformas permiten estructurar las publicaciones utilizando estándares de metada como Dublin Core [4], que está compuesto por 14 valores de datos que permiten que otras aplicaciones puedan entender de qué forma está estructurado un documento y de que se trata. Estas características son útiles para directorios, catálogos de revistas como también para Plataformas de contenido de Investigación como Google Scholar o Google Académico.

Google Scholar (GS) es una plataforma web que contiene los perfiles de investigadores de Universidades del Mundo, como las publicaciones que estas personas han realizados para conferencias, congresos, libros, revistas o patentes, además muestra el número de **citaciones** que tiene cada publicación, este valor representa el impacto que tiene esta publicación en el ámbito de Investigación.[5]

Además, el perfil muestra el valor del índice H, o hindex que permite cuantificar la producción de la investigación midiendo la productividad y el impacto del autor, esta es una medida internacional también usada por otras redes como Web of Sciene, este valor indica que un investigador con un índice H de 37, tiene 37 artículos citados al menos 37 veces [6]. Esta información sería útil poder extraerla y analizarla, la cual permitiría medir o evaluar la producción científica no solo de los investigadores sino de una Universidad o país, lamentablemente en la plataforma GS no existe forma de poder descargar la información de cada perfil, ni afiliación.

Con el objetivo de extraer estos datos de GS, investigamos diferentes métodos de extraer información de un sitio web utilizando minería de datos, específicamente, minería web, que contiene una técnica de extracción de datos llamado “Web Scraping”, con esta técnica y diferentes métodos web, logramos extraer los de GS, donde se realizó una comparación de la velocidad de extracción de datos y el formato de salida para conocer cual método permitía la forma más rápida, eficiente y estructurada de extraer los datos. En las pruebas que realizamos ningún método resulto ser óptimo al extraer datos de debido a que el proceso fue semi-automática, por lo que decidimos crear un algoritmo de Web Scraping utilizando el lenguaje R utilizando para crear patrón personalizado y facilitar el proceso de extracción de datos.

II. CONCEPTOS

A. GOOGLE SCHOLAR

Google Scholar (GS) es un buscador de Google lanzado al público en noviembre de 2004, orientado al ámbito académico y de investigación, posee una base de datos en Internet con los perfiles y publicaciones de investigadores de cerca de 4400 Universidades a Nivel Mundial, según listado del Ranking de Webometric del 2016 [7]. Las citaciones como los índices de impacto están incluidos

como indicadores de del área de ALMETRICS, que busca también proponer otros indicadores como número de visitas, número de descargas de las publicaciones científicas en un Repositorio. [8]. Se incluye un conjunto de trabajos de investigación científica de acceso abierto en diversas disciplinas, con documentos de revistas, congresos, tesis, libros, patentes, escritos por investigadores o académicos en diferentes idiomas. [9].

B. MINERIA DE CONTENIDOS

La minería de contenido o minería de texto intenta recopilar información significativa a partir del texto del lenguaje natural, extrayendo información de forma automática de un documento o página web, generalmente una gran cantidad de recursos textuales que no están estructurados, identificando patrones para leer o extraer los datos que implica el descubrimiento de nueva información. [10]

Las principales formas de minería de contenido web son:

- Minería de Datos no estructurada
- Extracción de datos estructurados
- Minería de datos semi-estructurada
- Extracción de datos multimedia

Dentro de la "Minería de Datos no estructurada" está la minería de documentos web y la minería de páginas web, que utiliza la técnica de web scraping para scanear datos. [11].

C. WEB SCRAPING

La Web Scraping (Raspado de páginas web), consiste en la extracción de una o varias páginas web de un sitio web que estén relacionadas mediante enlaces, para su manipulación, procesar parte o todo contenido para buscar patrones y extraer los datos para su análisis posterior "Fig. 1". Esta técnica es utilizada por los motores de búsqueda para extraer datos, también es considerada como el proceso de automatizar el método de copiar y pegar. [12]

Fig. 1 Esquema de la Vinculación del Web Scraping con la Web

Para hacer Web scraping es necesario analizar aspectos cómo:

1. Accesibilidad de los datos de origen.
2. Análisis de patrones de los datos.
3. Diferenciar la minería de datos web de otras opciones de minería
4. Frecuencia de Extracción de los datos.
5. Si existen otras alternativas de extracción de datos

D. MINERÍA DE TEXTO VS OTROS CONCEPTOS

Existen diferentes áreas que están vinculadas a la minería de textos, lo cual complementan está técnica, sin embargo, hay algunas diferencias entre estás.

- En la minería de texto, el texto es libre, no está estructurado.
- En la Minería de datos, los patrones se extraen del texto del lenguaje natural en lugar de las bases de datos.
- En la minería Web, las fuentes web están estructuradas por medio del lenguaje html, pero los datos no están estructurados.
- Recuperación de información, no se encuentra información genuinamente nueva, la información deseada simplemente coexiste con otras informaciones válidas "Fig. 2".

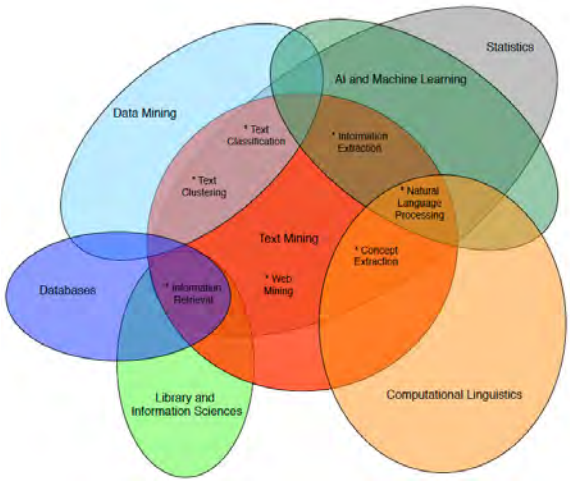

Fig. 2 Diagrama de la vinculación que existe entre Minería de Datos, Minería de Texto y Minería web, Extracción de Información.

E. LENGUAJE R

R, es un lenguaje de programación de código abierto, desarrollado actualmente por un grupo llamado Core Team. [13]. Es un lenguaje de script lo que no requiere ser compilado para ser ejecutado. Tiene similitud con otros lenguajes como C o C++, pero su mayor virtud es que mezcla diferentes características de otros lenguajes y paradigmas de programación. Dentro de las características de este lenguaje es que su orientación es para hacer minería y análisis de datos por lo que está compuesto por diferentes librerías o paquetes para realizar estas funciones, en este artículo se mencionan los paquetes utilizados para la realización del algoritmo.

III. TRABAJOS RELACIONADOS DE SCRAPER DE DATOS DE GS EN LA UTP

En marzo del 2016, debido a la necesidad de extraer los datos de Google Scholar del perfil de la Universidad (UTP) se realizaron varias pruebas de web scraping utilizando 4 métodos que se muestran en la "Tabla 1. Ha excepción del método de copiar y pegar, fue posible exportar los datos en formato .CSV, pero, no sin antes

realizar un proceso de depuración de los datos debido a que estos están unidos a otros textos que no eran de interés.

TABLA 1
MÉTODO DE WEB SCRAPING, APLICACIÓN A UTILIZAR Y FACILIDAD DE USO DEL MÉTODO

Métodos	Aplicación	Descarga	Conocimientos del usuario	Facilidadde uso
Copiar y Pegar	Manual	ninguno	ninguno	Fácil
Web scraping[14] Local Browser	Extensión Scraper Chrome	Gratuito	Técnico mínimo	Fácil
Web scraping Local Software[15]	Fminer	Pago (Trial)	Técnico intermedio	No es Fácil
Web Scraping Online[16]	Import.io	Pago (Free versión)	Técnico mínimo	Fácil

Para realizar pruebas con los Métodos de Web Scraping seleccionamos 5 perfiles en Google Scholar de Universidades [17]: *Universidad Fransisco Marroquin (UFM), Escuela Superior Politécnica del Litoral (ESPOL), Universidade Regional de Blumenau (FURB), Universidad Tecnológica de Panamá (UTP), Universidad de La Habana (UH).* El número de perfiles y publicaciones de universidades se muestran en la "Tabla 2".

TABLA 2
PERFILES DE UNIVERSIDADES SELECCIONADAS PARA WEB SCRAPING EN GOGGLE SCHOLAR

Universidad	País	#Perfiles	#Publicaciones
UFM	Guatemala	14	393
ESPOL	Ecuador	67	1061
FURB	Brasil	38	1360
UTP	Panamá	77	1434
UH	Cuba	79	2758

Para cada perfil de las Universidades seleccionadas se aplicó cada método, se extrajeron todas las publicaciones de los perfiles y se midió el tiempo de scraper. Según los resultados mostrados en la "Tabla 3", el método de **Web Scraping Online** resultó con un promedio de tiempo de extracción de datos de **7 minutos por perfil y 124 minutos por las publicaciones**, este método estructuró los datos al ejecutarlo, por lo que no hay un proceso de depuración.

TABLA 3
RESULTADO DEL TIEMPO PROMEDIO DE SCRAPER POR MÉTODO, DE LOS PERFILES Y PUBLICACIONES DE LAS 5 UNIVERSIDADES EN GS

	TIEMPO POR MÉTODO WEB SCRAPING Perfiles / Publicaciones (minutos)			
Universidad	**Copiar / Pegar**	**Local Browser**	**Local Software**	**Online**
UFM	8 / 130	2 / 35	3 / 50	2 / 35
ESPOL	42 / 354	9 / 95	14 / 140	9 / 94
FURB	24 / 445	5 / 122	8 / 179	5 / 120
UTP	50 / 482	11/129	17 / 189	10 / 127
UH	51 / 920	11 / 245	17 / 363	10 / 244
Promedio	**35 / 466**	**8 / 125**	**12 / 184**	**7 / 124**

Las pruebas dieron como resultado que el tiempo total para extraer los datos de una Universidad con un promedio de 35 perfiles y 466 publicaciones es de **2 horas 18 minutos**, considerando que no se extrajeron los detalles de cada publicación, el tiempo de extracción puede aumentar, también, es de considerar que el proceso fue semi-automático, por lo que se hizo necesario buscar otra alternativa para scrapear los datos, la alternativa planteada fue crear un algoritmo, que permitiera extraer los datos más rápido, completos y de forma estructurada, como también seleccionar un lenguaje que tuviera la facilidad de hacer análisis de datos, por eso se seleccionó el lenguaje R el cual incluía paquetes para scrapear páginas web.

IV. TRABAJOS RELACIONADOS CON EL LENGUAJE R PARA EXTRACCIÓN DE LOS DATOS DE GS

En las investigaciones que realizamos, se encontraron dos alternativas para extraer datos de Googles Scholar, una función y un paquete.

A. FUNCIÓN EN R "GScholarScraper"

Es una función en R creada en el 2012 por Kay Cichini, esta función permite Scrapear los perfiles y detalles de las publicaciones de un perfil determinado en Google Scholar.

Resultados de las pruebas de esta función:

- Solo permite extraer un perfil a la vez.
- No muestra a que perfil pertenecen las publicaciones extraídas, ni a que afiliación.
- No muestra a que afiliación pertenece el perfil.
- Es necesario colocar el año que se desea extraer las publicaciones del perfil, si no, se obtienen 0 datos.
- La entrada para hacer el Scraper es la URL del perfil, pero, no indica cual es el formato para introducir.
- Se obtuvieron errores al introducir la URL de algunos perfiles de las Universidades utilizadas.

- Esta es la tercera revisión de la Función y hasta noviembre de 2016 no tiene actualización, [18]

B. PAQUETE EN R "SCHOLAR"

Es un paquete en R, que proporciona funciones para extraer datos de Google Scholar, incluyendo, perfiles, citaciones, publicaciones y predicción del h-index. Fue creado por James Keirsted.en noviembre del 2015 y su última actualización es de junio de 2016. [19]

Resultados de las pruebas de este paquete:

- No tuvimos error al ejecutar las funciones del paquete.
- Se utilizó la función *get_profile()* para extraer los datos del perfil, se debe introducir el ID_user de GS, por lo que cada perfil se extrae por separado.
- Se utilizó la función *get_publications()*, para extraer los datos de detalles de las publicaciones hay que colocar el ID_USER.
- No indica a que usuario de GS pertenece los detalles de las publicaciones extraída.
- No muestra a que afiliación pertenece el perfil.

De las dos alternativas la opción que decidimos utilizar como componente del algoritmo fue el paquete en "Scholar" ya que permitía realizar la extracción de las publicaciones de un perfil, conociendo el ID del user.

IV. METODOLOGÍA PARA LA CREACIÓN DEL ALGORITMO

A. RECURSOS

- Aplicación R commander
- Aplicación R studio para Windows
- Paquete Rrvest, para leer todo el contenido HTML de una página web.
- Paquete Scholar, para extraer los detalles de las publicaciones de un perfil en GS
- Dependencias de los paquetes: xml2, plyr, wordcloud, dplyr, plot, qplot, string, gplot2.
- Computador con Windows 7 de 64 Bits, Dual Core de 2.2 GHz, y Memoria RAM de 3 GB.
- Computador con Windows 7 de 64 Bits, i7 de 3.40 GHz, y Memoria RAM de 8 GB.
- La velocidad de Internet en periodo de pruebas fue de 1.45 Mb de descarga y 1.90 de Carga.
- El periodo de desarrollo de los algoritmos fue de 6 meses, incluyendo mejoras por resultados de las pruebas.
- El periodo de prueba del algoritmo fue de 6 meses, incluyendo pruebas de velocidad, extracción de perfiles, extracción de publicaciones, pruebas de los métodos y comparación de algoritmos.

B. ANÁLISIS DE LA ESTRUCTURA DE DATOS EN GS

Se Analizaron los datos que queríamos copiar identificando patrones repetitivos en los códigos de la página web de Google Scholar, de tal forma que los fragmentos que había que Scrapear tuvieran la misma estructura y resultara más fácil crear ciclos de repetición. En la "Fig. 3", se muestra los perfiles de Afiliación de GS de la Universidad Tecnológica de Panamá donde se muestran bloques de contenidos similares por cada perfil.

Fig. 3 Listado de Perfiles de Google Scholar, afiliación Universidad Tecnológica de Panamá

Para poder realizar Web Scraping es importante conocer algunos elementos claves relacionados con una página web, esto es necesario para extraer los fragmentos HTML que se desean Scrapear:

1. DOM: es el Document Object Model, el código HTML de una página web es una jerarquía de Nodos anidados, y cada nodo contiene una etiqueta HTML, dentro de cada etiqueta puede haber otros nodos que contienen el "dato" que nos interesa identificar para extraer en formato de texto o número.
2. HTML: Es el lenguaje base de una página web, también llamada capa de contenido, el lenguaje html está compuesto por una serie de etiquetas html como: div, span, h1, aside, article, que se deben conocer para identificar el bloque HTML que se va a extraer.
3. CSS: Las "Hojas de estilo en cascada" o capa de presentación, es una serie de clases que se insertarn en el html por medio del atributo "class", esta permite darle un aspecto visual diferente a la página, algunas veces los datos dentro del código html estarán dentro de una clase CSS, por lo que es importante saber identificarlas.

En la "Fig. 4" se muestra el código HTML extraído del perfil de GS donde se puede ver las etiquetas:
<div class="gsc_1usr ***gs_scl"***> encierran los datos que nos interesan de este perfil (negritas).

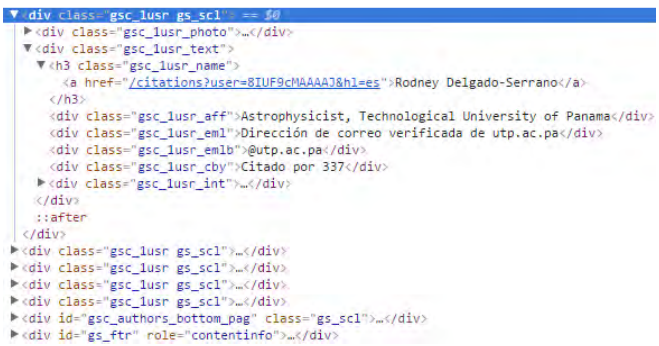

Fig. 4 Estructura HTML de Bloque de Perfil Scrapeado

C. BÚSQUEDA DE PATRONES

Analizamos el código HTML extraído de cada bloque de perfil en GS para buscar si los códigos html que contienen los datos tenían el mismo patrón y esquema de datos. Separamos cada uno de los nodos HTML "Tabla. 4", que contenían datos en bloques individuales que serán almacenados en variables para luego agruparlas en una tabla en R llamada data.frame, esta permitía almacenar diferentes tipos de datos.

TABLA 4
SCRAPER DE DATOS POR VALOR HTML, LOS RESULTADOS QUE SE OBTIENEN DE CADA BLOQUE Y LAS VARIABLES ASIGNADAS

Variable	**Valor HTML**	**Resultado**
url_perfil	read_html(url_GS)	Código HTML completo de primera página
Afiliación	html_text(,url_perfil, **h2.gsc_authors_header**)	Universidad Tecnológica
Perfil	html_node(url_perfil, **div.gs_scl**)	Código HML de perfil GS
Nombre	html_text(Perfil, **h3>a**)	Elida Obaldía
Url_perfil	html_atttr(Perfil, **href**)	https://scholar.google.es/citations?user=l8gpxI4AAAAJ&hl=es
Id_user	extraer_cadena(Url_perfil)	l8gpxI4AAAAJ

Al analizar cada bloque scrapeado de los perfiles de la primera página encontramos que existe una clase HTML que enmarca el contenido de cada perfil, esta clase *div.gs_scl* "Fig. 6" es un nodo que se repite, al utilizar realizar las pruebas con el paquete Rvest y la función *html_nodes(url_afiliacion,"div.gs_scl"*) con el parámetros de la clase identificada, R mostraró los bloques de contenido extraído que cumplían con este patrón html dentro del código, que en total fueron 10 nodos de perfil por cada página.

D. ANÁLISIS DE LOS DE LOS PERFILES

Cada afiliación de una Universidad u Organización en GS está compuesta por el listado de perfiles, cada perfil contenía la URL con el ID_USER. A su vez cada perfil integraba el listado de publicaciones y cada publicación contenía detalles de esa publicación. Nuestro objetivo fue desarrollar un algoritmo en dos procesos, extracción de perfiles y extracción de publicaciones de forma dinámica.

E. ESQUEMA DE LOS ALGORITMOS EN R

1. Algoritmo para Srapear Listado de Perfiles en GS

Se desarrolló un algoritmo para scrapear todos los ID_USER de los perfiles de una afiliación. El dato de entrada del algoritmo fue la URL de afiliación de una Universidad en GS. Se extrajo el enlace de cada perfil y se guardó de forma independiente. Con el enlace de cada perfil se realizó un ciclo de repetición, para extraer los campos que contenían: Nombre, afiliación, palabras claves, citaciones, y vincular la URL del perfil y el ID del perfil. Los datos fueron almacenados en un tipo de dato llamado data.frame, esta tabla temporal almacenaba los datos hasta que terminara el ciclo de repetición, al finalizar el ciclo, todos los datos se guardaron en un data.frame en R, que se podía accesarse y visualizar posteriormente "Fig.5"..

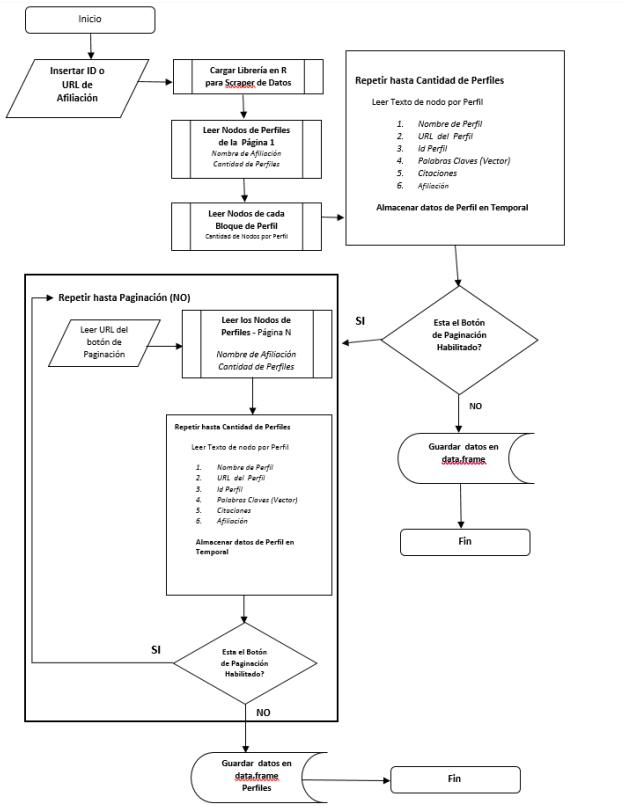

Fig. 5 Esquema de Algoritmo para Scrapear listado de Perfiles de una Afiliación en Google Scholar

2. Algoritmo para Srapear detalles de Perfiles

Este algoritmo utiliza la URL de cada perfil los cuales están almacenadas en un dataframe que fue creado en el algoritmo de listado de perfiles. Se realiza una lectura de la cantidad de perfiles guardados y se lee el campo URL que contiene el enlace de cada perfil de esta afiliación. Los nuevos datos de cada perfil campos que se extrajeron fueron: palabras claves, citaciones, citaciones 2011, hindex, hindex_2011, estos datos se vincularon a la URL del perfil y el ID del perfil. En este algoritmo se extrajo nuevamente las palabras claves, pero, pero cada palabra fue almacenada separado por coma para su posterior análisis. Los datos fueron almacenados en data.frame temporal hasta finalizar el ciclo de repetición "Fig.6".

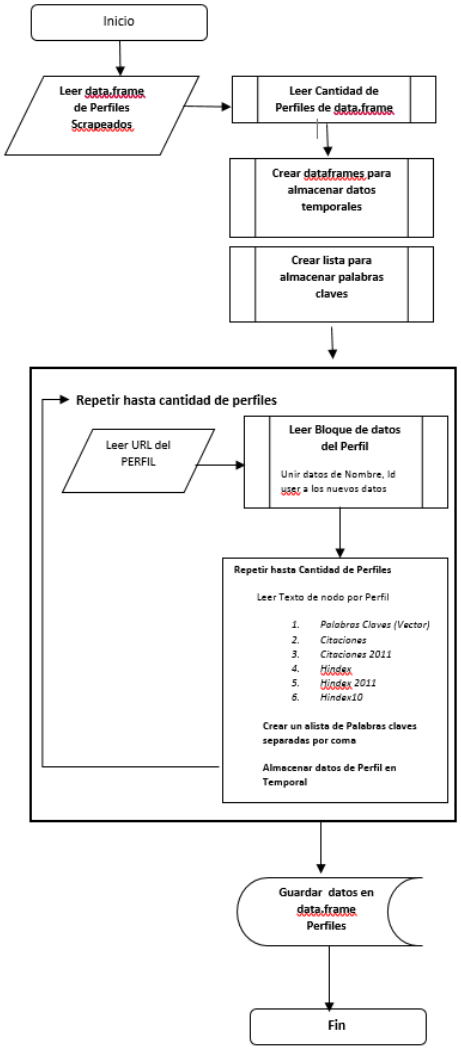

Fig. 6 Esquema de Algoritmo para Scrapear de detalles de los perfiles de una Afiliación en Google Scholar

3. Algoritmo para Srapear Publicaciones de Perfiles en GS

Se creó un segundo algoritmo para scrapear todas las publicaciones por cada perfil, en este algoritmo se utilizó el paquete (scholar) y específicamente la función *get_publications()* que permitió extraer las publicaciones y los detalles de cada. El algoritmo utilizó la tabla creada en el algoritmo 1, donde leía la URL de cada perfil para poder extraer el ID_USER y el nombres de cada perfil, el cual se vinculaba con las publicaciones extraídas La estructura de los detalles de las publicaciones se hizo en un ciclo de repetición hasta extraer todas las publicaciones por perfil y luego terminar el ciclo de la lectura de los id de los perfiles. Las publicaciones tenían estructuras diferentes ya que algunas publicaciones, eran de revistas, congresos, libros "Fig. 7 ".

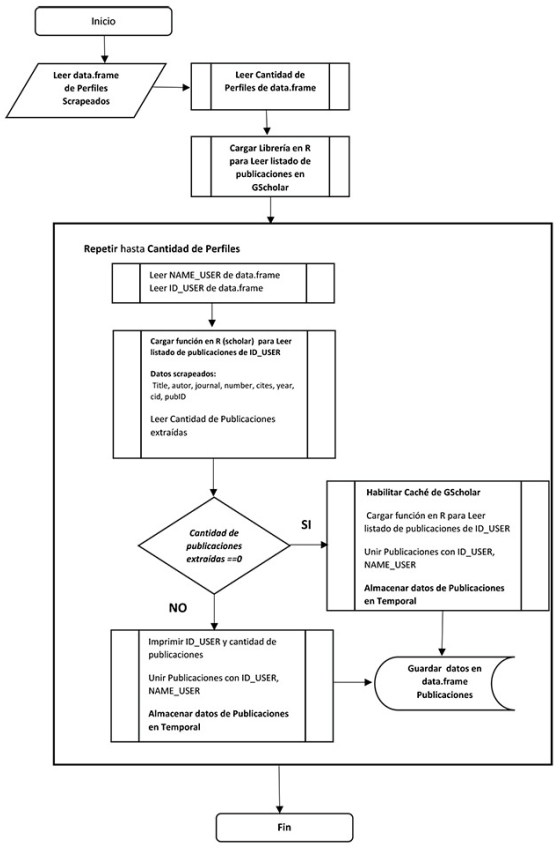

Fig. 7 Esquema de Algoritmo para Scrapear todas las Publicaciones por perfil de una Afiliación en Google Scholar

V. RESULTADOS

A. Pruebas y comparación de métodos y algoritmo en R

Prueba 1

Para realizar las primeras pruebas del algoritmo y validar el funcionamiento del algoritmo se seleccionaron 17 Universidades de forma aleatoria de una lista de 4400 afiliaciones en GS ordenadas por citación según Ranking de Webometrics "Tabla 5". La extracción de datos de los perfiles de estas Universidad tenía como objetivo validar diferentes elementos de los perfiles y publicaciones para poder mejorar el algoritmo. Se realizaron varias pruebas con cada perfil Y dependiendo de los errores que fueron encontrando se hicieron mejoras a las rutinas del algoritmo.

Listado de elementos que se mejoraron luego de la prueba 1:

1. Validar si en la afiliación no hubiera perfiles.
2. Validar que en la afiliación solo existiera una página por lo que el ciclo de repetición debiera terminar.
3. Validar que la afiliación tuviera un enlace de paginación activado, pero sin contenido en la siguiente página.

4. Validar que el número de perfiles por página fuera dinámico, ya que el usuario puede cambiar cuantos perfiles puede ver por página.
5. Validar que en el perfil no tenga publicaciones
6. Validar que el perfil no tenga palabras claves
7. Validar que en el perfil no tenga datos de citaciones, no hindex, hindex11.
8. Validar que las publicaciones en los perfiles tuvieran URL de detalles.
9. Validar que las publicaciones tienen diferentes estructuras según el contenido.
10. Validar que el nombre del perfil se agrupa con las publicaciones de ese perfil
11. Optimización de código, eliminando asignación de variables innecesarias, validación de respuestas antes de entrar a ciclos de repetición.

TABLA 5
LISTADO DE UNIVERSIDADES SELECCIONADAS PARA VALIDAR PRIMERA VERSIÓN DE ALGORITMO EN R

Universidad	País	Número de citaciones en GS
University of Illinois at Urbana Champaign	Estados Unidos	556164
University of Edinburgh	Reino Unido	439812
Universidad de Osaka	Japón	203376
Universidad Nacional Autónoma	México	146336
Universidad Complutense de Madrid	España	128024
Universidad de Chile	Chile	103997
Universidad Politécnica de Valencia	España	93120
Universidad de Costa Rica	Costa Rica	40068
Université de Franche Comté	Francia	37253
Universidad de Antioquia	Colombia	27410
Universidad de la República	Uruguay	26483
Universidad de La Habana	Cuba	8767
Universidad Nacional Costa Rica	Costa Rica	6747
Universidade Regional de Blumenau	Brazil	4032
Escuela Superior Politécnica del Litoral	Ecuador	3369
Instituto Tecnológico de Costa Rica	Costa Rica	2114
Fransisco Marroquin	Guatemala	1393

Las características de estos perfiles, había Universidades con una sola página de perfil, había páginas con más de 100 perfiles, perfiles con una sola publicación, perfiles sin palabras claves, sin hindex, sin detalles en las publicaciones, algún perfil tenía 3000 publicaciones.

Prueba 2

Para evaluar las correcciones del Algoritmo en R y la velocidad de extracción de datos, se realizó una comparación del primer algoritmo llamado "0 algoritmo R" y la versión optimizada y corregida del algoritmo "1 algoritmo R" utilizando los datos de las 5 Universidades iniciales (UFM), (FURB), (ESPOL), (UTP), (UH) con 55 perfiles y 1400 publicaciones. En la" Tabla 6" se muestra que el tiempo de scraper de los perfiles de estas Universidades utilizando el Método llamado "1 algoritmo en R" fue de 1.16 segundos por perfil vs los 14.87 segundos por perfil que demoró el primer algoritmo, este resultado es el promedio de 30 pruebas realizadas. Los tiempos de los resultados variaron en función de la velocidad de Internet en el momento de la prueba, por lo que se muestra un promedio del tiempo en la variable mean, el paquete y la función utilizada en R para la comparación fue "microbenchmark", el cual permite hacer comparación de varias funciones a la vez y un número de veces determinado.

TABLA 6
PRUEBA DE TIEMPO DE SCRAPER DE DATOS DE PERFILES GS UTILIZANDO ALGORITMO EN R

expr	min	lq (25%)	mean	uq (75%)	median	max	# Pruebas
1Algoritmo en R	1.0368	1.1238	**1.162**	1.2107	1.2256	1.2405	30
0Algoritmo en R	8.0121	8.0121	**14.866**	14.8663	21.7204	21.7204	30

Prueba 3

En la segunda prueba utilizamos 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. Según la "tabla 6" el tiempo promedio de extracción de datos fue de 4 minutos de Scraper. Esta prueba se incluyó 40% más perfiles y 60% más publicaciones, el resultado fue que solo se incrementó en 1 minuto el tiempo de extracción de los datos. En las pruebas realizadas con ambos métodos el tiempo de Scraper de los perfiles es de un minuto, también inferior al mejor promedio de los métodos que es de 7 minutos "Fig. 7". El resultado de los datos de perfiles fue almacenado en un tipo de dato en R llamado data.frame que permite exportarse al formato abierto .CSV para su uso en otra aplicación.

TABLA 7
PRUEBA DE TIEMPO PROMEDIO DE SCRAPER DE DATOS DE GS UTILIZANDO DIFERENTES MÉTODOS DE WEB SCRAPING

		Tiempo Scraper (minutos)			
Método	**#Perfiles / #Publicaciones**	**Perfiles**	**Publicaciones**	**Total**	**Horas**
Local (Copiar/Pegar)	55 / 1400	35	466	**501**	**8,21**
Local Browser	55 / 1400	8	125	**133**	**2,13**
Local Software	55 / 1400	12	184	**196**	**3,16**
Online	55 / 1400	7	124	**131**	**2,11**
1 Algoritmo en R	55 / 1400	1	2	**3**	**0,03**
2 Algoritmo en R	76 / 2232	1	3	**4**	**0,04**

Prueba 4

Se realizó una tercera prueba con el método "Algoritmo en R" seleccionando 15 Universidades con perfil en GS, 5 de ellas se habían utilizado en las pruebas anteriores. Las 10 nuevas universidades seleccionadas fueron: 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 Autonoma de Mexico (UNAM), Universidad de Osaka (OSAKAU), University of Edinburgh (UED), Universidad Politécnica de Valencia (UPV), University of Illinois at Urbana-Champaign (UILLINOIS), estas universidades tenían como mínimo 180 perfiles y 6500 publicaciones . En la prueba se Scrapearon todos los perfiles y todas las publicaciones de cada perfil incluyendo los detalles de cada publicación.

Los resultados de las pruebas en la "Tabla 8" muestran detalles por Universidad, donde el tiempo promedio de extracción de datos para las Universidades con menos de 100 perfiles y 3400 publicaciones fue de **2 minutos.**

El total de datos extraídos de las 15 Universidades fue de 8364 perfiles y 175,086 publicaciones donde el tiempo total fue de **122 minutos (2 horas 2 minutos)**, inferior al tiempo de cualquier método.

TABLA 8
CANTIDAD DE PUBLICACIONES POR UNIVERSIDAD EN GS Y TIEMPO DE SCRAPER CON ALGORITMO EN R

Universidad	País	#Perfiles	#Publicaciones	Tiempo (minutos)
UFM	Guatemala	14	393	**1**
ESPOL	Ecuador	67	1061	**1**
FURB	Brasil	38	1360	**1**
UTP	Panamá	77	1434	**1**
UH	Cuba	79	2758	**3**
UDELAR	Uruguay	182	6388	**2**
UCR	Costa Rica	230	6952	**2**
UFC	Francia	119	7063	**2**
UDEA	Colombia	383	8429	**3**
UCHILE	Chile	566	11433	**5**
UNAM	México	1329	12670	**11**
OSAKAU	Japon	460	13038	**4**
UED	Escocia	1471	14091	**14**
UPV	España	794	29835	**11**
UILLINOIS	Estados Unidos	2555	58181	**62**
		8364	**175086**	**122**

Prueba 5
El uso del paquete Scholar en el algoritmo de detalles de publicaciones nos permitió agilizar el desarrollo, sin embargo, al verificar los perfiles y publicaciones de estas Universidad en GS de forma manual, encontramos que el listado de las publicaciones no se extrajo de forma completa.

En las primeras 9 Universidades donde los perfiles tenían menos de 100 publicaciones, la cantidad de publicaciones extraías fue el total correcto. En las otras 6 Universidades algunos perfiles tenían como total de publicación 0 ó 100, al verificar los perfiles de forma manual, todos tenían publicaciones y en algunos casos tenían hasta 3000 publicaciones. Según "Tabla 9" la cantidad de perfiles con datos incorrectos fueron: en la UNAM de 566 Perfiles 167 estaban correctos, en OSAKAU de los 1329, solo 138 estaban correctos, de la UED, 140 perfiles de 460 estaban correctos, de la UPV, 387 de 794 y de la ULLINOIS 2250 de 2555.

TABLA 9
PERFILES EXTRAIDOS POR UNIVERSIDAD EN GS Y PERFILES CON CANTIDAD DE PUBLICACIONES ERRONEAS

Universidad	#Perfiles	#Publicaciones	Perfiles extraídos con datos
UFM	14	393	14
ESPOL	67	1061	67
FURB	38	1360	38
UTP	77	1434	77
UH	79	2758	79
UDELAR	182	6388	182
UCR	230	6952	230
UFC	119	7063	119
UDEA	383	8429	383
UCHILE	**566**	11433	**137**
UNAM	**1329**	12670	**138**
OSAKAU	**460**	13038	**140**
UED	**1471**	14091	**147**
UPV	**794**	29835	**387**
UILLINOIS	**2555**	58181	**2250**

El problema que encontramos es que la función get_publications() que extrae los detalles de las publicaciones, contiene una variable (FLUSH=false) que scrapea lo que GS tiene en caché, cuando se habilitó a (FLUSH=true), algunos perfiles donde solo se extrajeron 100 publicaciones, cambiaron a total de publicaciones de 2000 ó 3000, algunos que tenían valor 0 se les cargó otro valor , pero, no el correcto, los perfiles que tenían algún valor , pasaron a tener 0 publicaciones, por lo que el valor de (FLUSH= true/false) no permite scrapear la cantidad de datos correctos cuyos perfiles tengan más de 100 publicaciones.

B. Funciones desarrolladas en el algoritmo para Visualizar datos extraídos de GS

En el algoritmo se crearon 6 funciones que utilizan un solo parámetro para extraer o estructurar los datos, aunque es necesario conocer R para utilizarlos, su uso es muy simple.

PubGS_research(url_afiliacion): función que permite extraer los Nombres y ID_user de una afiliación en Google Scholar, url_afiliacion es la URL en GS de la Universidad, los datos pueden ser asignados a un data.frame para almacenarlos.

PubGS_hindex(Id_users), función que extrae los detalles de cada perfil de la afiliación, leyendo los UD_user del data.frame extraído en la función anterior "Fig. 8"

	Nombre	word_key	url_user	Id_user	citaciones	cita_2011	hindex	hindex_2011
1	Reinhardt Pinzon Adames		/citations?user=1TCsmUAAAAJ&hl=es&oe=ASCII	1TCsmUAAAAJ	841	426	10	6
2	Elda de Obaldia	Material Science, Renewable Energy, Diamond thin...	/citations?user=I8gpx4AAAAJ&hl=es&oe=ASCII	I8gpx4AAAAJ	795	315	8	4
3	Oscar M. Ramirez	Structural and Earthquake Engineering	/citations?user=Ko4JMoMAAAAJ&hl=es&oe=ASCII	Ko4JMoMAAAAJ	383	246	6	6
4	Rodney Delgado-Serrano	Astrophysics, Astronomy, Physics, Mathematics, ...	/citations?user=8IUF9cMAAAAJ&hl=es&oe=ASCII	8IUF9cMAAAAJ	326	237	5	5
5	Adan Vega Saenz	soldadura, elementos finitos, mecanica computaci...	/citations?user=Dfi8fvYAAAAJ&hl=es&oe=ASCII	Dfi8fvYAAAAJ	312	167	10	8
6	Gilberto Axel Chang, PhD	Structural Engineering, Seismic Engineering	/citations?user=KGGMPHgAAAAJ&hl=es&oe=ASCII	KGGMPHgAAAAJ	302	182	3	3
7	Ildeman Abrego Castillo	Ciencias de los Materiales	/citations?user=74B2jbkAAAAJ&hl=es&oe=ASCII	74B2jbkAAAAJ	299	82	5	3
8	Héctor Montes Franceschi	Robotica, Control	/citations?user=EpT2FMcAAAAJ&hl=es&oe=ASCII	EpT2FMcAAAAJ	293	175	10	8
9	Vladimir Villarreal	Ambient Intelligence, Ubiquitous Computing, Ambi...	/citations?user=52pqkbwAAAAJ&hl=es&oe=ASCII	52pqkbwAAAAJ	248	205	9	8
10	JOSE R. FABREGA-DUQUE	tropical hydrology, environmental chemistry, hydr...	/citations?user=RiE3jucAAAAJ&hl=es&oe=ASCII	RiE3jucAAAAJ	236	86	7	5
11	Norma L. Miller	Mapas conceptuales, pedagogias activas, aprendi...	/citations?user=uxyQt2AAAAAJ&hl=es&oe=ASCII	uxyQt2AAAAAJ	195	113	8	7
12	Eleicer Ching Prado	Ciencia de materiales	/citations?user=EKdDC4AAAAJ&hl=es&oe=ASCII	EKdDC4AAAAJ	167	37	6	4
13	Humberto Rodriguez	control of robots, image processing, control of UA...	/citations?user=d1F8qX8AAAAJ&hl=es&oe=ASCII	d1F8qX8AAAAJ	151	41	9	5
14	Pablo Montero	food technology, shelf life, food processing	/citations?user=y1yykoAAAAJ&hl=es&oe=ASCII	y1yykoAAAAJ	125	125	3	3
15	Ramiro Vargas	Structural Engineering, Seismic Design, Steel Struc...	/citations?user=Br4QpIMAAAAJ&hl=es&oe=ASCII	Br4QpIMAAAAJ	113	83	6	5
16	Casilda Saavedra	climate change, environmental sciences	/citations?user=Ul4EB_8AAAAJ&hl=es&oe=ASCII	Ul4EB_8AAAAJ	101	87	2	2
17	Maritza Morales Batista	inteligencia artificial, open data, linked data, big d...	/citations?user=UdcSsuYAAAAJ&hl=es&oe=ASCII	UdcSsuYAAAAJ	80	32	5	3
18	Carlos Vergara-Chen	Marine ecology, molecular ecology	/citations?user=r23s60QAAAAJ&hl=es&oe=ASCII	r23s60QAAAAJ	79	79	5	5
19	David Vega	hidrologia	/citations?user=sNTQYqEAAAAJ&hl=es&oe=ASCII	sNTQYqEAAAAJ	79	34	5	4
20	Alexis Mojica	Geofisica Aplicada	/citations?user=6VyYFAsAAAAJ&hl=es&oe=ASCII	6VyYFAsAAAAJ	62	47	5	4
21	Yessica Saez	Smart Grid, Probabilidad y Procesos Aleatorios, S...	/citations?user=68YDp6EAAAAJ&hl=es&oe=ASCII	68YDp6EAAAAJ	62	62	5	5

Fig. 8 Vista de Data frame de datos Scrapeado de los perfiles de la UTP en GS

PubGS_publications(Id_users), función que extraer todas las publicaciones y sus detalles y las une al perfil e ID de cada usuario "Fig. 9".

	title	author	journal	number	cites	year
1	Introduction to active thermochemical tables: Sev...	B Ruscic, RE Pinzon, ML Morton, G von Laszevski, ...	The Journal of Physical Chemistry A	108 (45), 9979-9997	275	2004
2	Active thermochemical tables: Accurate enthalpy o...	B Ruscic, RE Pinzon, ML Morton, NK Srinivasan, MC...	The Journal of Physical Chemistry A	110 (21), 6592-6601	205	2006
3	Active Thermochemical Tables: thermochemistry f...	B Ruscic, RE Pinzon, G Von Laszewski, D Kodeboyin...	Journal of Physics: Conference Series	16 (1), 561	137	2005
4	Metadata in the collaboratory for multi-scale chem...	C Pancerella, J Hewson, W Koegler, D Leahy, M Lee...	International Conference on Dublin Core and Meta...	121-129	68	2003
5	Comparison of different three-site interaction pote...	E Guardia, R Pinzon, J Casulleras, M Orozco, FJ Luq...	Molecular Simulation	26 (4), 287-306	43	2001
6	A collaborative informatics infrastructure for multi...	JD Myers, TC Allison, S Bittner, B Didier, M Frenkla...	Challenges of Large Applications in Distributed En...	2004. CLADE ...	31	2004
7	A collaborative informatics infrastructure for multi...	JD Myers, TC Allison, S Bittner, B Didier, M Frenkla...	Cluster Computing	8 (4), 243-253	28	2005
8	On the solvation shell of Na+ and Cl- ions in aceto...	E Guardia, R Pinzon	Journal of Molecular Liquids	85 (1), 33-44	24	2000
9	A grid service-based active thermochemical table ...	G Von Laszewski, B Ruscic, P Wagstrom, S Krishna...	International Workshop on Grid Computing	25-38	15	2002
10	Implantation and damage under oblique low-energ...	R Pinzón, HM Urbassek	Physical Review B	63 (19), 195319	10	2001
11	Hydroclimate projections for Panama in the late 2...	J Fábrega, T Nakaegawa, R Pinzón, K Nakayama, O...	Hydrological Research Letters	7 (2), 23-29	2	2013
12	Estimates of Biomass and fixed carbon at a Rainf...	R Pinzón, J Fábrega, D Vega, EN Vallester, R Abrego...	Air, Soil and Water Research	79	2	2012
13	Increase in simple precipitation intensity index in ...	K Nakayama, C Botta, E Vallester, R Pinzon, , Fabr...	[illegible]	81 [illegible] 68 (4), I_16...	2	2012
14	FURTHER REFINEMENTS OF THE BOND DISSOCIATIO...	B Ruscic, RE Pinzon, ML Morton, B Wang, AF Wagne...	Ohio State University		2	2003
15	Study of Seasonal Rainfall infiltration Via Time-Lap...	A Mojica, I Díaz, AC Ho, F Ogden, R Pinzón, J Fábre...	Air, Soil and Water Research	133-139	1	2013
16	Rainfall-CN (Curve Number) relationships in a trop...	D Fábrega, R José, R Pinzón, E Vallester, D Vega	Revista Facultad de Ingeniería Universidad de Ant...	170-176	1	2012
17	Collaborating for Multi-Scale Chemical Science	WH Green	Massachusetts Institute of Technology		1	2006
18	ACTIVE THERMOCHEMICAL TABLES: ON THE ACCUR...	B Ruscic, RE Pinzon, A Fernandez, ML Morton, D Ko...	Ohio State University		1	2006
19	Sandia National Laboratories, Livermore, CA 9455...	JD Myers, TC Allison, S Bittner, B Didier, M Frenkla...	University of California, Berkeley	94720	1	2004
20	REFINEMENTS OF THE BOND DISSOCIATION ENERG...	B Ruscic, RE Pinzon, ML Morton, AG Császár, JF St...	Ohio State University		1	2004

Fig. 9 Datos Scrapeados de todas las Publicaciones por perfil en GS

PubGS_totalCitas(data_publicaciones), muestra una tabla con el número de citaciones y publicaciones por año.

	year	total_citas	total_publi
2	1978	0	10
3	1980	16	8
4	1981	11	7
5	1982	0	1
6	1983	1	3
7	1984	0	1
8	1987	5	5
9	1988	15	3
10	1989	0	1
11	1990	1	3
12	1991	1	2
13	1992	8	8
14	1993	32	8
15	1994	364	21
16	1995	158	8
17	1996	120	16
18	1997	66	13
19	1998	133	10
20	1999	213	13

Fig. 10 Listado de citaciones y publicaciones por año

PubGS_GraficaCitas(citas_anio), crea una gráfica que compara el número de citaciones por año, y el número de publicaciones, también muestra los valores más altos de cada variable "Fig. 11"

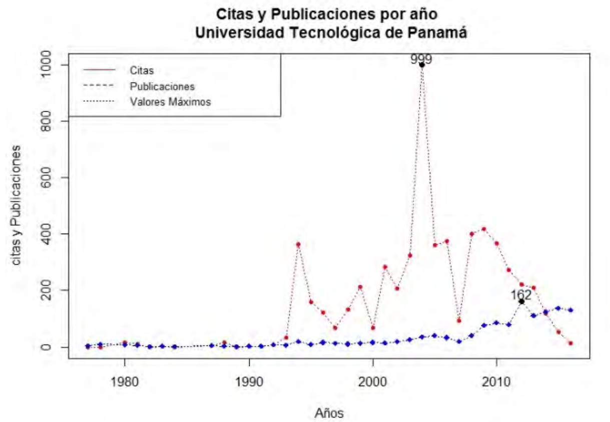

Fig. 11 Número de Perfiles y Publicaciones vs Tiempo de Scraping

PubGS_wordKey(data_publicaciones): función que genera gráfica de barra del listado de palabras claves según perfiles Scrapeados. "Fig. 12".

Fig. 12 Nube de Palabras claves más frecuentes en perfiles

PubGS_wordKey_freciencia(data_publicaciones): función que genera frecuencia de palabras claves mención según perfiles Scrapeados . "Fig. 13".

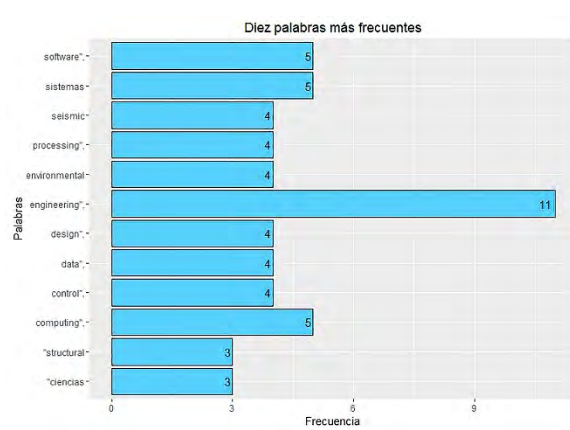

Fig. 13 Palabras claves más frecuentes usadas en los Perfiles, en forma de barras o nubes de palabras

PubGS_journal(data_publicaciones), Muestra el listado de Journals y el número de publicaciones y citaciones de cada uno, analizando los datos Scrapeados. "Fig. 14".

	journal	num_citaciones	num_publi
545	RIDTEC	0	2
532	Revista Indexada I+D Tecnológico: RIDTEC	0	1
523	GTIDEE	0	1
440	RIDTEC Revista de I+ D Tecnológico	0	1
76	LA FORMACIÓN PRÁCTICA EN LA UNIVERSIDAD Y SU...	0	1
349	Prisma Tecnológico	8	71
82	Prisma	0	7
355	Revista Prisma Tecnológico	2	4
514	Prisma Tecnológico. ISSN	0	2
513	Prisma Tecnológico, ISSN	0	1
485	Prisma Tecnológicao	0	1
458	Prisma Tecnológico | Vol.	0	1
351	Prisma Técnológico	0	1
350	Prisma Tecnólogico	0	1

Fig. 14 Listado de Journals y número de publicaciones

VII. CONCLUSIÓN

La minería de texto es una herramienta que permite extraer datos estructurada o no estructurada y convertirla en información de interés para el usuario. En este artículo hemos realizado diferentes pruebas que muestran que la técnica de Web Scripting es una opción a tomar en cuenta al momento de extraer datos de cualquier sitio web. Las pruebas realizadas utilizando diferentes métodos muestran que, aunque la técnica es buena no es funcional para todos los propósitos ya que no se logró el objetivo inicial de extraer los datos de GS, debido a que el proceso resulto semi-automático.

Implementar un algoritmo, aunque más compleja a la hora de desarrollarlo debido al conocimiento que se debe tener acerca del Lenguaje resultó ser mejor al extraer los datos con una estructura personalizas en un tiempo reducido en comparación con cualquiera de los otros métodos utilizado, a su vez el algoritmo permitió generar datos que no necesitan depurarse para ser usados.

Con el resultado positivo de las pruebas utilizando el algoritmo en R, se mejora el tiempo de extracción, considerando elementos que mejorar, pero los resultados son de gran beneficio para las Universidades y las personas involucradas en la medición del impacto de la producción científica y académica, porque tendrán una herramienta para minimizar el trabajo de extracción lo que les permite analizar esta información de forma más rápida y oportuna.

VI. TRABAJOS FUTUROS

Se ha logrado crear un algoritmo para realizar pruebas para extraer los datos de los detalles de las publicaciones, eliminando el uso del paquete "Scholar" y extraer las publicaciones completas y detalles de las publicaciones que serán integradas a una versión del algoritmo.

Se está trabajando en la creación de un algoritmo que permita extraer los detalles de cada artículo incluyendo las URL donde están alojadas estas investigaciones para ser vinculadas con los detalles de las publicaciones y los perfiles extraídos.

Los detalles de los perfiles, hindex y citaciones serán vinculados a la plataforma de investigadores de la UTP llamada SICUTP, para mostrar este valor en el perfil de cada investigador, además se está trabajando en el componente para poder utilizar los detalles de publicaciones extraídas.

Enlace de Algoritmo Versión 2.0 utilizando el paquete "Scholar"
https://bitbucket.org/dannymu/ejemplos-de-r/src

BIBLIOGRAFÍA

[1] Buenas prácticas en la construccion de una identidad academica online para una universidad, 2014, Enrique Orduña-Malea, Emilio Delgado López-Cózar
[2] Tim Berners-Lee: "El papel no desaparecerá, siempre habrá cosas que nos guste leer en ese formato", http://www.lne.es/asturama/2012/02/15/tim-berners-lee-papel-desaparecera-habra-cosas-guste-leer-formato/1199452.html.
[3] Berners-Lee and M. Fischetti. Weaving the Web. HarperOne, San Francisco, USA, 1999.
[4] Evolución del Dublin Core Metadata Initiative , José A. Senso, Antonio de la Rosa Piñero
[5] J. M. Russell, M. J. Madera Jaramillo, and S. Ainsworth, "El análisis de redes en el estudio de la colaboración científica," Redes. Rev. Hisp. para el Análisis Redes Soc., vol. 17, no. 2, pp. 39–47, 2009.J
[6] Measuring your research impact: H-Index, 2017, http://guides.library.cornell.edu/c.php?g=32272&p=203391
[7] Methodology, 2016, http://www.webometrics.info/en/Methodology
[8] Alonzo-Arévalo, M. Vázquez Vásquez, Altmetrics y alfabetización científica, vol. 12, no. 12 pp. 14-29, 2016
[9] Silva Ayçaguer, Luis Carlos. (2012). El índice-H y Google Académico: una simbiosis cienciométrica inclusiva. ACIMED, 23(3), 308-322. Recuperado en 06 de diciembre de 2016
[10] D. McDonald, U. Kelly, and JISC - Joint Information Systems Committee, "The Value and Benefit of Text Mining to UK Further and Higher Education," JISC Digit., no. March, pp. 1–32, 2012.B. Ian H. Witten, Text Mining, pp. 2-3
[11] Ghosh Dastidar, D. Banerjee, S. Sengupta, An Intelligent Survey of Personalized Information Retrieval using Web Scraper, pp. 24-31, 2016
[12] Alternativas para realizar web scraping, Mayo 2016, http://felicianoborrego.com/alternativas-para-realizar-web-scraping/
[13] Richard Cotton, Learning R, O´RELLY, pp. 3-4, 2013
[14] Scraper extensions for Google Chrome, https://chrome.google.com/webstore/detail/scraper/mbigbapnjcgaffohmbkdlecaccepngjd?utm_source=chrome-app-launcher-info-dialog
[15] Fminer, http://www.fminer.com/
[16] Import.io, https://www.import.io/
[17] Listado de Perfiles de Universidades a Nivel Mundial en Google Scholar, 2016, http://www.webometrics.info/en/node/169
[18] GScholarScraper_3.1.R. ,2012,https://github.com/gimoya/theBioBucket-Archives/blob/master/R/Functions/GScholarScraper_3.1.R
[19] Package 'scholar', 2015, https://cran.r-project.org/web/packages/scholar/index.html
[20] Simon Munzert, Christian Rubba, Peter Meißner, ominic Nyhuis, Automated Data Collection with R. 2015 pp.19-25
[21] R. Baron, T. Baldwin, D. Martinez, Web Scraping Made Simple with SiteScraper, 2009