Expresiones Regulares para el Tratamiento de la Privacidad en Navegadores Web

Donovan Riaño Enriquez
Facultad de Ingeniería
Universidad Nacional Autónoma de México
Ciudad Universitaria, México
donovanriano@gmail.com

Guillermo Molero-Castillo
Facultad de Ingeniería
Universidad Nacional Autónoma de México
Ciudad Universitaria, México
gmoleroca@fi-b.unam.com

Alejandro Velázquez-Mena
Facultad de Ingeniería
Universidad Nacional Autónoma de México
Ciudad Universitaria, México
mena@fi-b.unam.mx

Everardo Bárcenas
Facultad de Ingeniería
Universidad Nacional Autónoma de México
Ciudad Universitaria, México
barcenas@fi-b.unam.com

***Resumen*—En este artículo se presenta el uso de expresiones regulares para el tratamiento de la privacidad en navegadores Web. En la actualidad, el uso de las expresiones regulares y el reconocimiento óptico de caracteres ha sido fundamental para múltiples implementaciones de Software, soluciones en temas de la industria, robótica, procesamiento de lenguaje natural, desarrollo de aplicaciones, entre otras. La prueba se hizo en cinco navegadores Web. Como resultado se logró la detección de anuncios publicitarios que distraen o roban la atención de los usuarios. A través de esta identificación se busca bloquear la publicidad en la Web y darle al usuario privacidad.**

***Index Terms*—Ad-Blockers, Expresiones Regulares, Navegadores Web, Publicidad, Reconocimiento Óptico.**

I. Introducción

En la actualidad, se vive en el periodo de la cuarta revolución industrial. Las primeras tres revoluciones estuvieron marcadas por [1]: las máquinas; la electricidad y la industria automovilística; y las computadoras, como tecnologías de información para automatizar la producción en general. En esta cuarta revolución industrial se tiene a la Inteligencia Artificial, Internet de las cosas, impresión en 4D, cómputo en la nube, entre otras, como áreas con gran potencialidad para nuevos desarrollos tecnológicos [2][3].

Hoy en día, la computación es un área poderosa en la resolución de problemas de la sociedad actual, relacionadas, por ejemplo, con la salud, educación, comunicaciones, entretenimiento, seguridad, comercio electrónico, procesos industriales y otros campos de conocimiento que demandan soluciones enfocadas en la satisfacción del usuario final.

No obstante, a lo largo de la historia han existido notables desarrollos científicos y tecnológicos, que hacen que en él presente se viva en una era digital que facilita, en cierta medida, las actividades diarias de las personas. Sin embargo, la tecnología trajo consigo también aspectos negativos, como el mal uso de la publicidad en los navegadores Web.

Así, los usuarios de Internet se han percatado que al navegar en diferentes sitios Web existe, cada vez más, una mayor cantidad de publicidad en el contenido. Es por esto qué se comenzó a buscar la manera de que estos anuncios "estorbosos" no distraigan o roben la atención de lo consultado en la Web [4].

En este trabajo se presenta el uso de expresiones regulares, apoyado por el reconocimiento óptico de caracteres (OCR) como opción para el tratamiento y bloqueo de la privacidad en distintos navegadores Web. Se hicieron pruebas en distintos navegadores, logrando detectar anuncios publicitarios que distraen la atención de los usuarios.

II. Antecedentes

Con base en el análisis del ciclo de vida de las tecnologías emergentes que presenta cada año Gartner, www.gartner.com/en, se observa la relación entre el tiempo y la visibilidad de tienen estas tecnologías en el contexto actual, soportado por sus avances, desinterés y consolidaciones en un determinado periodo, esto es, a corto, mediano y largo plazo. Estas tecnologías buscan mejorar la vida cotidiana de las personas y de la sociedad en general.

Una de estas tecnologías es el posicionamiento de Internet y los dispositivos inteligentes, que trajeron consigo un cambio en la forma de comunicación [5]. A su vez, estos cambios traen consigo un nuevo campo, conocido como Internet de la cosas, en el que se busca que los objetos físicos estén perfectamente integrados en la red de información y puedan convertirse en participantes activos en los distintos procesos de negocio.

Es en estos procesos, a través de Internet, se añaden otros servicios para interactuar con los "objetos inteligentes". Sin embargo, se tiene el problema del manejo de la seguridad y privacidad [6].

Cisco muestra, a través de sus investigaciones, los datos crecientes sobre el routing de Internet en intervalos semestrales [7]. Con base a estas investigaciones, se estima que cada año

crece la cantidad de dispositivos conectados a Internet a nivel mundial. En la Fig. 1 se muestra el número de la población mundial y los dispositivos conectados por persona. Se observa que en esta última década se tienen más dispositivos conectados que personas.

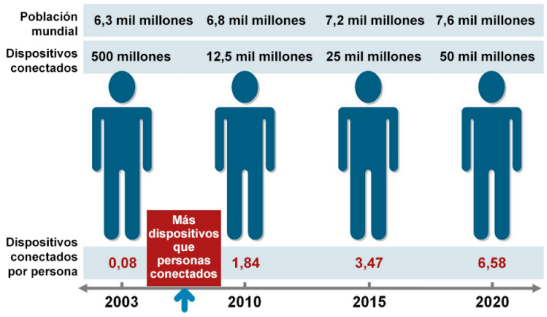

Figura 1. Cisco IBSG. Fuente [7].

Por otro lado, dado el potencial y las estrategias del marketing online actual [8], esta actividad hace que sea utilizada, con mayor frecuencia, como parte importante de las campañas de fidelización de las marcas, puesto que es un canal de comunicación que consigue la interacción entre el cliente potencial y la marca. Entre las acciones enfocadas por el marketing online destacan [8]:

- Fidelizar al cliente.
- Incrementar la imagen de la marca y sus ventas.
- Generar promociones y pruebas del producto.
- Incentivar la repetición de compra del producto.
- Realizar una campaña de comunicación directa y personalizada.

Precisamente, una forma de tratar la publicidad dirigida y la privacidad en navegadores Web es a través de Expresiones Regulares (Regex, RegExp o RE). Estas expresiones son una forma de representar cadenas de caracteres que se ajustan a cierto patrón [9]. Además, son un mecanismo flexible y eficiente para el procesamiento de textos [10]. Sus aplicaciones son diversas, por ejemplo, validación de campos de formularios, identificación de cadenas de texto en redes sociales, comandos de búsqueda, entre otras [11][12].

II-A. Manejo de la privacidad de los usuarios

En la actualidad, el manejo de la información y la privacidad está en una etapa crítica [13]. Empresas como Facebook, Twitter, Google, Amazon, entre otras, han incluido, en sus sitios Web, aplicaciones o motores de búsqueda para el acaparamiento de la información. Estas empresas proporcionan "gratuidad" por el servicio, pero utilizan la información de los usuarios, a conveniencia, para fines publicitarios. Entre los pilares para tener una adecuada seguridad de la información destacan [13]:

- Autenticación. Para identificar la entidad comunicante y la fuente de datos.
- Control de acceso. Para prevenir el uso no autorizado de los recursos.
- Confidencialidad. Para proteger los datos contra la revelación no autorizada.
- Integridad. Para garantizar la no alteración o destrucción de los datos de manera no autorizada.
- No repudio. Para dar prueba del origen de los datos o de entrega de los mismos.
- Disponibilidad. Para garantizar la continuidad de la accesibilidad y utilización por las entidades autorizadas.

Estos servicios se proporcionan mediante mecanismos de seguridad solos o combinados, como: cifrado, firma digital, mecanismos para el control del acceso, mecanismos para la integridad de datos, intercambio de autenticación, relleno de tráfico, control de encaminamiento, entre otras acciones.

Con respecto a la publicidad en los navegadores Web, ésta cada vez va en aumento, poniendo en riesgo la privacidad de los usuarios. Esto es, al hacer búsquedas en la Web de cierto tipo de compras o productos, hoy en día hay saturación de información en el sitio Web donde se está posicionado. El problema no sólo influye en el manejo de la información, sino en la lentitud del flujo de información por cuestiones de la publicidad.

Otro aspecto importante en las búsquedas de los usuarios es el manejo, almacenamiento e interpretación de la gran cantidad de datos recolectados. Por lo que, estos cúmulos de datos, bajo un estricto tratamiento, también pudieran aportar información útil para un mejor manejo de la publicidad dirigida [14].

II-B. Ad-blockers de los Navegadores Web

En los últimos años, se han incrementado los esfuerzos por implementar bloqueadores de publicidad Web [15]. Estos inicialmente se implementaron para el navegador *Firefox*, el cual a largo de los años fue mejorando hasta lograr un navegador sin publicidad. Pero las empresas anunciantes buscaron la forma de seguir mandando publicidad. Por lo que, en poco tiempo se empezaron a crear controversias [4]. Controversias por pérdidas importantes de posibles ingresos. Por lo que, se hicieron también mejoras en los bloqueadores de publicidad, conocidos actualmente como Ad-blockers [16].

Otro término que en la actualidad también se utiliza son los Trackers en el marketing, que son indicadores y rastreadores de efectividad de campañas publicitarias dirigidas. Estas herramientas van guardando información por medio de las *cookies*, y éstas a su vez proporcionan la ubicación de las búsquedas realizadas.

De esa manera, al ir creciendo la demanda y las posibilidades de bloquear la publicidad, se fueron implementando nuevos Ad-blockers en la mayoría de los navegadores Web, así como en las aplicaciones para dispositivos móviles. En este tipo de dispositivos la publicidad ha encontrado el camino de realizar anuncios sobre las búsquedas de los usuarios.

Un dato importante a considerar es que, durante estos últimos años, Google controla el 85 % del negocio mundial de publicidad en buscadores y cerca del 50 % de toda la publicidad online [17]. Las personas y la comunidad en general ve a Google como un servicio [18], pero tras eso hay

una tecnología que contiene funciones específicas e incluye extensiones exclusivas y restrictivas.

En este sentido, la publicidad y las compras por Internet son cada vez más demandadas y rápidas, por ejemplo, las compras en línea ahorran tiempo y distancia. Sin embargo, el uso de los sitios Web para este tipo de compras sirve a las empresas para aprender de las experiencias y necesidades de los usuarios. Estos se traducen como patrones de comportamiento en diferentes extractos sociales, ya sea de manera local o regional.

III. Testing de publicidad en los navegadores Web

Con base en los antecedentes mencionados, el algoritmo propuesto en este trabajo fue implementado en Python, asistido por las librerías Tesseract [19], Pillow y OpenCV [20], así como el paquete Tesseract-Ocr. En el código siguiente se muestra un extracto del algoritmo implementado, donde la primera expresión regular utilizada fue "oferta" con sus diferentes variantes, esto es, mayúsculas, minúsculas y una combinación de éstas.

```
import cv2
import re
from PIL import Image
import pytesseract

img = cv2.imread(r'sears.jpg')
print(pytesseract.image_to_string(img))

print ("\n\nImpresión de la imagen\n")
print ("Se muestra la lista de la información\n\n")

with open ('Textos/sears.txt','r') as texto:
        cad = [linea.split() for linea in texto]
        lista = str(cad)

patronOferta = re.findall (r'\b oferta|Oferta|OFERTA|ofertas|Ofertas|OFERTAS \b',lista)
print("\nLa prueba para la palabra oferta/ofertas es:")
print (patronOferta)
print("Se ha encontrado {} veces".format(len(patronOferta)))
```

Es importante destacar que la efectividad de las Expresiones Regulares depende de la lectura de Tesseract. Así, las pruebas se realizaron en navegadores Web de escritorio. Estos navegadores fueron: Opera, Google Chrome, Mozilla Firefox, Safari y Microsoft Edge, de los cuales se obtuvieron datos antes, durante y después de las navegaciones. De igual manera, se recolectaron datos de la navegación con la inclusión de Ad-Blockers.

En el caso de Google Chrome, de acuerdo con [21] es un caso especial, pues se ha demostrado que a pesar de borrar el historial de navegación, su vector de búsquedas (Pagerank) va almacenando información de sus usuarios para generar publicidad dirigida [22][23]. Esta publicidad dirigida es su principal fuente de ingreso. La prueba de publicidad en los navegadores Web, a través de Expresiones Regulares, se validaron a través de las siguientes palabras y símbolos:

- Producto / Oferta
- Descuento / %
- Precio / $
- Rebaja / Promoción
- Especial / Bajo
- Venta / Liquidación
- Ahorro / Internet
- Marcas: Nike©, Adidas©, Amazon©, Huawei©, Liverpool©, Nissan©, Lenovo©, Intel©, Samsung©.

Se consideraron las palabras y símbolos mencionados, puesto que comúnmente se emplean para anuncios publicitarios [24]. Además, dado que la publicidad no sólo juega con los aspectos visuales, sino también con las letras, tamaños y estilos, sobre estas palabras se amplió el rango de búsqueda con plurales, mayúsculas, minúsculas, acentos, y primera letra en mayúscula y demás minúsculas.

IV. Resultados

Con base al testing realizado, las Fig. 2 y 3 muestran los resultados obtenidos de publicidad después de una serie de búsquedas específicas en el navegador Opera y Google Chrome sin Ad-blockers, respectivamente.

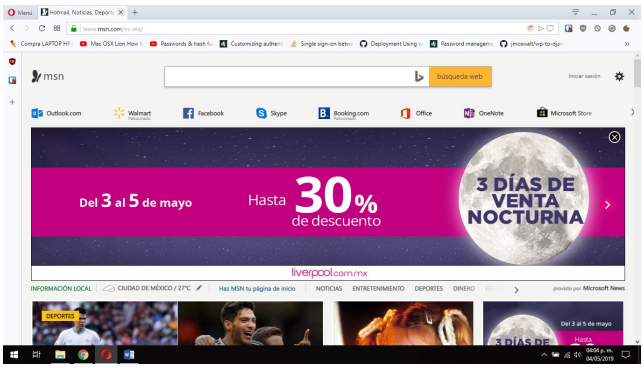

Figura 2. Identificación de publicidad después de una serie de búsquedas en Opera sin Ad-blockers.

En el caso del navegador Opera se observó en gran medida la invasión de publicidad de la empresa Liverpool. Mientras que Google Chrome fue invadido de publicidad de Sears.

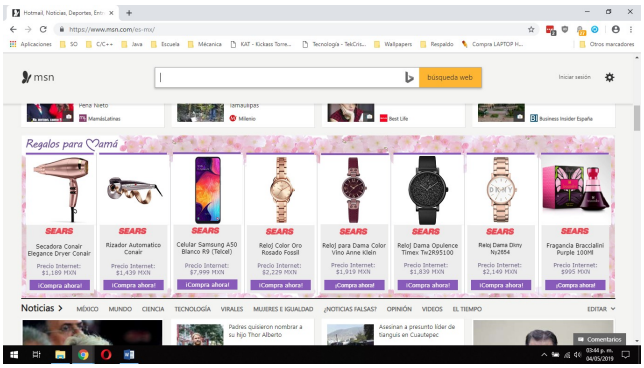

Figura 3. Identificación de publicidad después de una serie de búsquedas en Google Chrome sin Ad-blockers.

Por otro lado, las Fig. 4 y 5 muestran los resultados del reconocimiento óptico de caracteres, y la verificación de palabras por medio de las Expresiones Regulares, respectivamente. Cabe Destacar que sólo se realizó una vez el proceso del reconocimiento óptico de caracteres por navegador Web. Para el reconocimiento óptico se utilizaron las herramientas Tesseract, Pillow y OpenCV, cuyos resultados fueron almacenados en archivos de texto plano. Posterior a esto, sobre estos archivos

de texto se hicieron las verificaciones de las Expresiones Regulares.

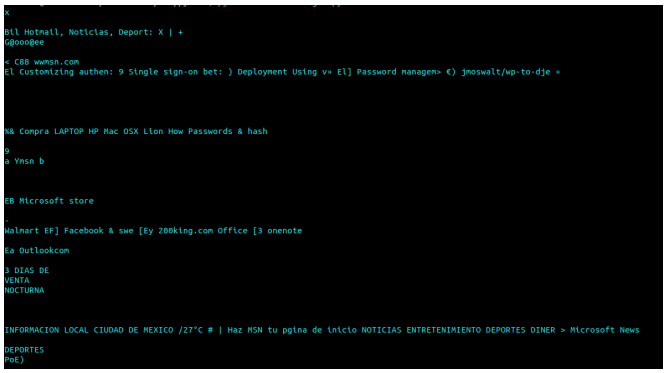

Figura 4. Reconocimiento óptico de caracteres a través de Tesseract, Pillow y OpenCV.

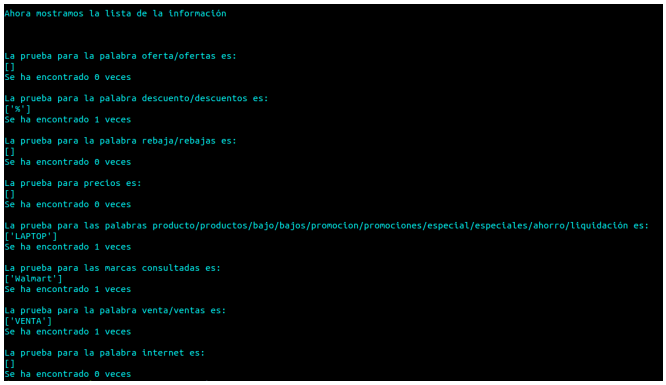

Figura 5. Verificación de coincidencias mediante Expresiones Regulares, posterior al reconocimiento óptico de caracteres.

Sobre la verificación de coincidencias a través de las Expresiones Regulares, se observó que las palabras con mayor invasión fueron: Venta, %, LAPTOP, Samsung, Walmart, SEARS e Internet. En la Tabla 1 se resume algunos ejemplos de publicidad detectada, cuatro en Opera y cuatro en Google Chrome.

Tabla I
RESULTADOS OBTENIDOS POR TESSERACT Y LAS EXPRESIONES REGULARES

Expresiones Regulares encontradas	
Navegador Web	Palabra
Opera	%
Opera	LAPTOP
Opera	Walmart©
Opera	Venta
Chrome	LAPTOP
Chrome	SEARS©
Chrome	Samsung©
Chrome	Internet

Por otro lado, se detectó en la mayoría de los navegadores Web, Walmart es la empresa que predominaba con mayor publicidad invasiva.

V. CONCLUSIONES

Los notables avances en la tecnología contemporánea traen consigo también consecuencias negativas para el usuario final, como la invasión de anuncios publicitarios en la Web. Anuncios que son dirigidos con base en las búsquedas, necesidades e intereses de los usuarios.

El uso de las Expresiones Regulares resultó útil, así como él del reconocimiento óptico de caracteres para la identificación de palabras persuasivas con anuncios publicitarios en los navegadores Web. Estos anuncios abarcan términos relacionados con venta e información de productos por Internet.

El uso de Ad-blockers refleja que la publicidad dirigida es un mecanismo que no se puede evitar, pero si moderar. La función de los bloqueadores de publicidad sólo es ocultar los anuncios, pero los algoritmos de Google logran cada vez inundar una mayor cantidad de publicidad dirigida.

Es importante destacar que recolectar información de los usuarios no es una mala práctica, sino que ésta debe ser para fines que beneficien a éstos. El inconveniente de tener publicidad estorbosa es que distraen la atención de los usuarios y se carezca de una adecuada seguridad de la información.

Para publicidad dirigida pudieran cubrirse otras opciones no distractoras, por ejemplo, a través de correos electrónicos, apartados específicos en los navegadores y aplicaciones especializadas, donde los usuarios puedan consultar ofertas publicitarias.

REFERENCIAS

[1] L. L. Heber. “Las Tres Revoluciones Industriales”. 2014. *El País*. [En línea]. Disponible en: https://elpais.com/elpais/2014/10/17/media/1413577081_550723.html.

[2] B. Bollig, J. P. Katoen, C. Kern, M. Leucker, D. Neider, D. R. Piegdon. “Libalf: The automata learning framework”. *International Conference on Computer Aided Verification*. 2010. [En línea]. Disponible en: https://link.springer.com/content/pdf/10.1007/978-3-642-14295-6_32.pdf.

[3] R. G. Herrera. “La inteligencia artificial: ¿Hacia dónde nos lleva?”. *¿Cómo ves?*. 1999. [En línea].

[4] F. Pezzino. “Ad blockers y publicidad en internet”. 2015. [En línea].

[5] C. Gálvez. “Aplicación de transductores de estado-finito a los procesos de unificación de términos”. *Ciência da Informação*, 35(3), 2006. [En línea]. Disponible en: http://www.scielo.br/pdf/ci/v35n3/v35n3a07.

[6] S. Haller, S. Karnouskos, and C. Schroth. “The internet of things in an enterprise context”. *Future Internet Symposium*,páginas 14–28, 2008. [En línea]. Disponible en: https://www.alexandria.unisg.ch/46642/1/fis2008-haller-final.pdf.

[7] D. Evans. “Internet de las cosas. Cómo la próxima evolución de Internet lo cambia todo”. *CiscoInternet Bussiness Solutions Group-IBSG* , 11(1):4–11, 2011. [En línea]. Disponible en: https://www.cisco.com/c/dam/global/es_mx/solutions/executive/assets/pdf/internet-of-things-iot-ibsg.pdf.

[8] CECARM. *“Marketing Online: Potencial y Estrategias”*. Murcia, España. 2019. [En línea]. Disponible en: https://www.cecarm.com/Guia_Marketing_Online_Potencial_y_Estrategias_-_CECARM.pdf-6120.

[9] R. Pomol-Acosta, C. González-Segura, S. González-Segura. “Una herramienta didáctica para el aprendizaje interactivo de expresiones regulares” 2013. [En línea]. Disponible en: http://repositorio.uigv.edu.pe/handle/20.500.11818/804.

[10] R. C. Beltran. “El uso de expresiones regulares en la detección de errores escritos: implicaciones para el diseño de un corrector gramatical”. *El valor de la diversidad (meta) lingüística: Actasdel VIII congreso de Lingüística General*. 2008. [En línea]. Disponible en: https://dialnet.unirioja.es/servlet/articulo?codigo=4007478.

[11] A. J. Gallego. “La jerarquía de chomsky y la facultad del lenguaje: consecuencias para la variación y la evolución”. *Teorema: Revista internacional de filosofía*,47–60, 2008. [En línea]. Disponible en: https://www.jstor.org/stable/43047450?seq=1#page_scan_tab_contents.

[12] I. García-Campos. “Herramienta para la corrección automática de autómatas finitos”. 2017. [En línea]. Disponible en: https://riull.ull.es/xmlui/handle/915/5846.

[13] J. A. Sanchéz, L. López, J. F. Martínez. “Solución para garantizar la privacidad en el internet de las cosas”. *El profesional de la información*. 24(1):62–70, 2015. [En línea]. Disponible en: http://oa.upm.es/41198/1/INVE_MEM_2015_227635.pdf.

[14] M. D. Ortiz, L. J. Aguilar, L. M. Marín. “Los desafíos del marketing en la era del big data”. *e-Ciencias de la Información*,1-31, 2016. [En línea]. Disponible en: https://www.scielo.sa.cr/pdf/eci/v6n1/1659-4142-eci-6-01-00016.pdf.

[15] P. Cerezo. “Ad blocking: el modelo publicitario digital, a revisión”. *Cuadernos de periodistas:revista de la Asociación de la Prensa de Madrid*, 81–89, 2016. [En línea]. Disponible en: http://www.cuadernosdeperiodistas.com/media/2016/11/Pepe-Cerezo.pdf.

[16] A. F. Londaitz. “Publicidad en los celulares: Publicidad invasiva vs. derecho a la privacidad”. *PhD thesis, Universidad del Salvador*. 2011. [En línea]. Disponible en: https://racimo.usal.edu.ar/4312/.

[17] Google. “Bienvenido a Google, la mejor empresa para trabajar”. Expansion.com. 2013.

[18] J. Jarvis. “Y Google, ¿cómo lo haría?”. *Madrid: Gestión*,2000. [En línea]. Disponible en: https://narrativabreve.com/2013/10/libro-google-jeff-harvis.html.

[19] K. Sajjad. “Automatic license plate recognition using python and opencv”. *Department of Computer Science and Engineering, MES*, College of Engineering, Kuttippuram, Kerala, 2010. [En línea]. Disponible en: https://pdfs.semanticscholar.org/bddf/1200eb17f239e4dce2a9cec938eb8cf305f5.pdf.

[20] C. Patel, A. Patel, and D. Patel. “Optical character recognition by open source ocr tooltesseract: A case study”. *International Journal of Computer Applications*, 55(10):50–56, 2012. [En línea]. Disponible en: http://www.assistivetechnology.vcu.edu/wp-content/uploads/sites/1864/2013/09/pxc3882784.pdf.

[21] L. Lombardero. “Trabajar en la era digital: tecnología y competencias para la transformación digital”. *LID Editorial*, 2015. [En línea]. Disponible en: https://bit.ly/2T2nshj.

[22] P. L. Guzmán. “Motores de búsqueda; soluciones con aplicaciones de Google”. *Revista e-FORMADORES*. [En línea]. Disponible en: http://red.ilce.edu.mx/sitios/revista/e_formadores_ver_11/articulos/paola_ver11.pdf.

[23] F. Pedroche. “Métodos de cálculo del vector pagerank”. *Bol. Soc. Esp. Mat. Apl*,39:7–30, 2007. [En línea]. Disponible en: http://personales.upv.es/~pedroche/inv/_docs/FPedrochev4(sema).pdf.

[24] M. Vallez. “Keyword Research: métodos y herramientas para identificar palabras clave”. *BiD: textos universitaris de biblioteconomia i documentació*, 27, 2011. [En línea]. Disponible en: http://eprints.rclis.org/16956/1/Keyword_Research_mvallez.pdf.