Revista Española de Documentación Científica
36(3), julio-septiembre 2013, e018
ISSN-L:0210-0614. doi: http://dx.doi.org/10.3989/redc.2013.3.956

ESTUDIOS / *RESEARCH STUDIES*

Evolución de los factores de posicionamiento web y adaptación de las herramientas de optimización

Jorge Morato*, Sonia Sánchez-Cuadrado*, Valentín Moreno*, Jose Antonio Moreiro*

*Universidad Carlos III de Madrid
Correo e: jmorato@inf.uc3m.es

Recibido: 13/03/2012; 2ª versión:13/06/2012; Aceptado: 18/06/2012

Cómo citar este artículo/ Citation: Morato, J.; Sánchez-Cuadrado, S.; Moreno, V.; Moreiro, J. A. (2013). Evolución de los factores de posicionamiento web y adaptación de las herramientas de optimización. Revista Española de Documentación Científica, 36(3):e018. doi: http://dx.doi.org/10.3989/redc.2013.3.956

Resumen: Las herramientas de optimización web, *Search Engine Optimization* (SEO), se utilizan para analizar y mejorar los sitios web en relación a distintos factores de posicionamiento. Esta investigación estudia la evolución de las estrategias de posicionamiento web y analiza la adaptación de las herramientas de optimización a estos factores. Además, se estudian qué factores de posicionamiento están presentes en las herramientas SEO más populares. En la fase experimental se analiza el grado en que las estrategias de optimización mejoran el posicionamiento, y en qué medida se encuentran esas funcionalidades en las herramientas SEO. Adicionalmente se han analizado foros y blogs oficiales para descubrir nuevas pautas de evolución de los motores de búsqueda y el grado en que las herramientas SEO pueden adaptarse a dichos cambios. Aunque estas herramientas optimizan el posicionamiento, los resultados sugieren la necesidad de introducir importantes mejoras que aumenten su potencialidad futura.

Palabras clave: Herramientas SEO; optimización en motores de búsqueda; posicionamiento web.

Evolution of web positioning factors and adaptation of optimization tools

Abstract: *Search Engine Optimization* (SEO) tools are designed to analyze and optimize resources regarding positioning factors. Web positioning techniques are applied in order to improve the relevancy of web resources. Webmasters usually use SEO tools to analyze a web site according to some positioning factor. They are required to be updated to achieve two basic goals: to increase the user's satisfaction while searching the web and to decrease web spamming. We have studied the trends that affect positioning algorithms and optimization techniques. Several SEO tools were analysed in order to learn which functionalities have been implemented. Furthermore, an experiment was performed to test how the positioning factors help optimize the results and if these factors are present in the functionalities found in the SEO tools. Finally, a literature review was carried out to detect future trends in search engines' algorithms. Results show that SEO tools help in the optimization process but to an insufficient degree; therefore the algorithm's evolution study suggests that there is a need for major updates in the short term.

Keywords: SEO tools; search engine optimization; web positioning.

EL POSICIONAMIENTO EN LOS BUSCADORES WEB

Los motores de búsqueda web son un campo ampliamente estudiado en el área de la Recuperación de Información (Baeza-Yates y Ribeiro-Neto, 2011; Manning y otros, 2008). Pero sin duda, uno de los temas más fascinantes son los algoritmos de posicionamiento (Marchiori, 1997; Brin y Page, 1998; Kobayashi y Takeda, 2000; Long y Suel, 2003; Moreno, 2005; Morato y otros, 2005). Estos algoritmos son los encargados de determinar el orden de los resultados de acuerdo a la relevancia de los documentos para cada una de las consultas que ejecuta el usuario.

A mediados de los 90 ya existían sistemas de recuperación web con gran popularidad. El crecimiento acelerado de la Web imposibilitó el mantenimiento de directorios validados manualmente como *Yahoo*, impulsando el rápido desarrollo de los motores de búsqueda. Cada buscador creó su propio motor y su particular algoritmo de posicionamiento. Sin embargo, los mecanismos de indización automatizada y estimación de relevancia que empleaban, convirtieron a los motores en herramientas fácilmente manipulables. En 1996, se alertaba del empleo de técnicas fraudulentas (*spamindexing* o *spamming web*) para modificar los resultados (Flynn, 1996). Y a pesar, de que buscadores como Altavista aplicaban un posicionamiento correcto en términos de precisión (Leighton, y Srivastava, 1999), la facilidad de su manipulación les hizo extremadamente vulnerables. Principalmente debido a que el algoritmo de posicionamiento se fundamentaba en factores presentes en la página web (Wall, 2011).

Un elemento clave en el progreso de los buscadores fue la valoración de los recursos mediante el análisis de enlaces externos de los usuarios web. Esta estrategia fue propuesta por Marchiori (1997) y Kleinberg (1998) e implementada posteriormente en el buscador *Google* mediante el algoritmo *Pagerank* (Brin y Page, 1998). Este algoritmo asigna mayor relevancia a aquellos recursos que reciben más enlaces. El algoritmo de *Google* valora que el texto del enlace que apunta al recurso coincida con los términos de la consulta, teniendo en cuenta el *Pagerank* de la página de procedencia. Ambos factores, número y texto de enlaces entrantes, son dos estrategias cuya complejidad para ser manipulados resulta difícil, ya que no dependen de los autores, sino de otros recursos externos a la página (Arbildi, 2005).

No obstante, estos factores tuvieron su respuesta en el *spamming web*. Por ejemplo, para manipular el recuento de enlaces se crean páginas web denominadas granjas de enlaces (*link-farms*). Su contenido principal son enlaces salientes a los recursos que se quiere posicionar. Otra técnica fraudulenta relacionada es el *bombing*; consiste en enlazar un recurso con un mismo texto predefinido desde diferentes sitios web y de forma masiva. El *bombing* usa como descriptor de la página web enlazada el texto del enlace entrante. La eficacia del *bombing* ha sido utilizada y comprobada con fines comerciales y difamatorios. Neutralizar estas técnicas ha requerido mucho esfuerzo a los buscadores. Desde el 2007, se han podido resolver diversos ejemplos de *bombings* de tipo difamatorio mediante la modificación, caso a caso, del algoritmo o eliminando resultados (Karch, 2011). Existen otras estrategias de *spamming web* como: ofrecer páginas falsas (*cloaking* y *doorway*), repetir palabras clave en la página para falsear recuentos (*stuffing*), ocultar texto en el código fuente o con el mismo color que el fondo de la página, incluir texto con menor tamaño del normal, o redirecciones de páginas web (Gyöngyi y Garcia-Molina, 2005).

Para combatir el impacto del *spam* se trabaja en detectar las técnicas fraudulentas. Se usan heurísticas, ponderación de factores, análisis de factores robustos a la inserción masiva de términos no relacionados con el contenido, cálculo de la densidad del término en lugar de frecuencia absoluta, etc. (Morato y otros. 2005). Los motores intentan minimizar el *spamindexing*, penalizando cualquier resultado que presente una característica sospechosa, incluso si no es el caso. Estas penalizaciones no son usualmente percibidas por el usuario que utiliza el buscador, debido al volumen de información duplicada en la Web. Estas estrategias de acción-reacción entre web *spammers* y buscadores web, y el modo en que los primeros superan los filtros de los segundos se registra periódicamente en las publicaciones sobre el tema. Un ejemplo es la actualización de *Google* en 2003 denominada Florida, en la que numerosas publicaciones reclamaban conocer su naturaleza (Thies, 2004). A esta actualización han seguido otras, normalmente descritas en los blogs de los buscadores. Concretamente, el 4 de febrero de 2011, en el blog de *Google* apareció la noticia de una actualización que afectaría al 12% de los resultados. Casi un mes más tarde, el 11 de marzo, se anunció la actualización de *Caffeine*, con un impacto previsto de al menos un cambio en el 35% de las páginas de resultados (Singhal, 2011). También *Microsoft* con un algoritmo inicialmente manipulable (Wall, 2006), se ha ido aproximando a los factores más relevantes según *Google*, hasta el punto de que éste ha llegado a denunciar las similitudes entre ambos (Wingfield, 2011).

Desde el punto de vista del diseño web, se utilizan herramientas SEO como técnicas de desarrollo para mejorar la visibilidad de un recurso en Internet optimizando su posición en el ranking de resultados. Cumplir las expectativas de los usuarios al devolver los resultados de una consulta es un elemento decisivo. Por eso, el diseño y el estudio de los algoritmos de posicionamiento se ha convertido en un campo crítico tanto para los motores web como para las herramientas SEO. La investigación sobre herramientas SEO estudia el grado de ajuste a esos algoritmos. Esto implica que motores y herramientas SEO compitan entre sí para seguir siendo eficaces ordenando los resultados. Por tanto, los algoritmos que aplican los motores son dinámicos, y evolucionan para mejorar su funcionamiento de cara al usuario, y también para evitar estrategias para su manipulación intencionada.

METODOLOGÍA

El objetivo de este trabajo es conocer el grado en que las herramientas utilizadas por los expertos en posicionamiento se acercan a los algoritmos internos de los motores web. Se ha elaborado un estudio para detectar los factores que los motores de búsqueda consideran más influyentes en el posicionamiento (Fase 1). Después se ha buscado la adecuación de las herramientas SEO a la mejora del posicionamiento de un sitio web (Fase 2). Por último se ha analizado el desarrollo de los algoritmos y si su evolución limita el uso futuro de las herramientas SEO (Fase 3).

Fase 1: Factores de posicionamiento en herramientas SEO

Para el análisis de los factores de posicionamiento en herramientas SEO partimos de la hipótesis de que una funcionalidad está implementada en una herramienta SEO si se considera pertinente para el posicionamiento. Se han seleccionado 21 herramientas SEO (Anexo I) elegidas de acuerdo a su popularidad (número de enlaces entrantes y opiniones en foros especializados), y disponibilidad para su evaluación (eliminación de servicios de pago y sin restricciones en el acceso). Luego se han identificado las funcionalidades y se han listado con una breve descripción (Tabla I).

Tabla I. Funcionalidades en herramientas SEO

Funcionalidad en SEO	Descripción
Alerta de problemas	Alerta de cambios en el posicionamiento o caídas del servidor.
Análisis enlaces	Estructura, diseño y atributos de los enlaces. Incluyendo origen del tráfico.
Análisis etiquetas meta	Estudio de las palabras clave en las etiquetas Meta.
Búsquedas en *search engine*	Búsqueda de la palabra clave en distintos motores de búsqueda.
Búsquedas palabras clave usuarios	Popularidad de las palabras clave para las búsquedas en un periodo de tiempo. Se sugieren también palabras y frases relacionadas.
Categorización	Métodos para otorgar una categoría a las páginas para ayudar a los motores de búsqueda a clasificar la pagina de forma correcta.
Chequeo de enlaces	Análisis de problemas de conectividad y comunicación de los enlaces.
Chequeo direcciones red	Chequeo de las direcciones que pretenden intercambiar enlaces.
Chequeo *Dmoz directory*	Comprobar si la página está incluida en el directorio de referencia DMOZ.
Comparación páginas primeras posiciones	Análisis de las páginas de los primeros puestos para obtener patrones de optimización.
Constructor de páginas	Herramienta de ayuda para la creación del código HTML de una página.
Contenido de la página	Relación entre texto de la página y el código fuente.
Creación páginas *doorway*	Posibilidad de editar páginas *doorway* para la optimización de la página.
Densidad de palabras clave	Análisis de la frecuencia y la densidad de aparición de las palabras clave.
Editor HTML	Ayuda a la edición de los campos HTML (p.e.etiquetas Alt)
Efectividad palabras clave	Grado de dificultad según las palabras clave en función de su popularidad como término de búsqueda y páginas con las que compite.
Envío automático de páginas	Envío automático de la página a los buscadores.
Envío de páginas manual	Envío manual de la página a los buscadores.
Formato palabras clave	Estudio de los formatos de las palabras clave para mejorar su valoración.
Ftp *upload*	Subida de archivos al servidor mediante FTP.
Gestión de dominios	Gestión de las licencias de los dominios de las páginas web.
Historial de cambios	Se registran los cambios de posición y otros factores de la página.
HTML validador	Chequeo de la corrección del código HTML.
Informes	Generación de informes de resultados, almacenados para su uso posterior.
Mercado de enlaces	Gestión de intercambio de enlaces con sitios relacionados por temática.
Pagerank	Ranking o estimación del mismo según el pagerank de *Google*.
Pagerank dc	Ranking de los enlaces de *Google* en cada uno de los *datacenters*.
Páginas indexadas	Análisis de las páginas indexadas de un sitio web por distintos motores de búsqueda (p.e. calculado mediante el operador *site*:)
Pago por inclusión	Similar PPC sin necesidad de posicionamiento orgánico con SEO.
Popularidad enlaces	Se mide el grado de popularidad de una página según el número y la procedencia de los enlaces que la apuntan.
PPC	*Pay Per Click*. Esponsorizado a cambio del cobro de click por enlace.
Programador	Automatización con un programador de tareas.
Ranking palabras clave	Cada palabra clave obtiene, para cada página web concreta, un ranking. El ranking es la posición del recurso para cada palabra en un motor.
Ranking tráfico	Ranking de la página según el tráfico que recibe.
Rastreo del tráfico	Rastreo del tráfico que recibe una página web: número de usuarios, páginas vistas, procedencia, etc.
Simulador de *spider*	Visión de la página web para los robots de indexación. También visualiza *snippets* como la etiqueta HTML Meta *Description* y *keywords*.
Tamaño de la página	Número de bytes que ocupa la página web.
Test de velocidad	Test que mide la velocidad de carga de la página.
Versiones de la página	Versiones anteriores en *Way Back Machine* (web.archive.org).

En la Figura 1 se muestran las 39 funcionalidades encontradas y el número de herramientas SEO que la implementan. La funcionalidad más extendida es la "búsqueda de palabras clave" presente en un 68,2% de las herramientas SEO. Las siguientes funcionalidades más relevantes son: "popularidad de enlaces" con 63,6% y "Ranking de palabras clave" con un 59,1%. Un análisis pormenorizado de la naturaleza de las funcionalidades muestra que varias tienen más orientación de servicio al *webmaster* que de mejora al posicionamiento. Por ejemplo: creación de informes, alta en buscadores, PPC, ayuda a la edición de páginas en HTML, carga de archivos mediante FTP, alertas, o automatización de tareas.

Observamos que las herramientas SEO se centran en los factores directos presentes en la página, porque son más fáciles de identificar. Los factores indirectos y los factores externos a la página presentan mayor dificultad para su análisis. Entre los factores indirectos que suelen ser considerados están la popularidad de una búsqueda, la corrección del código y el tiempo de carga de la página. De los factores externos a la página encontrados destacan el análisis de la competencia y el número de enlaces entrantes. Desde el punto de vista de las herramientas SEO permiten realizar una comparativa según el número de funcionalidades que incluyen. Algunas de estas herramientas cubren un amplio rango de funcionalidades, mientras otras son específicas de una o unas pocas funcionalidades. Según el número de funcionalidades destacan las herramientas *SEOpen Toolbar* (28 funcionalidades), *Web CEO Version 6.0* (con 21), y *AddWeb™ Website Promoter 8* (20). El valor medio en número de funcionalidades es 9,2 con una mediana de 6 (σ 7,4).

Fase 2: Adecuación de las herramientas SEO en la mejora del posicionamiento

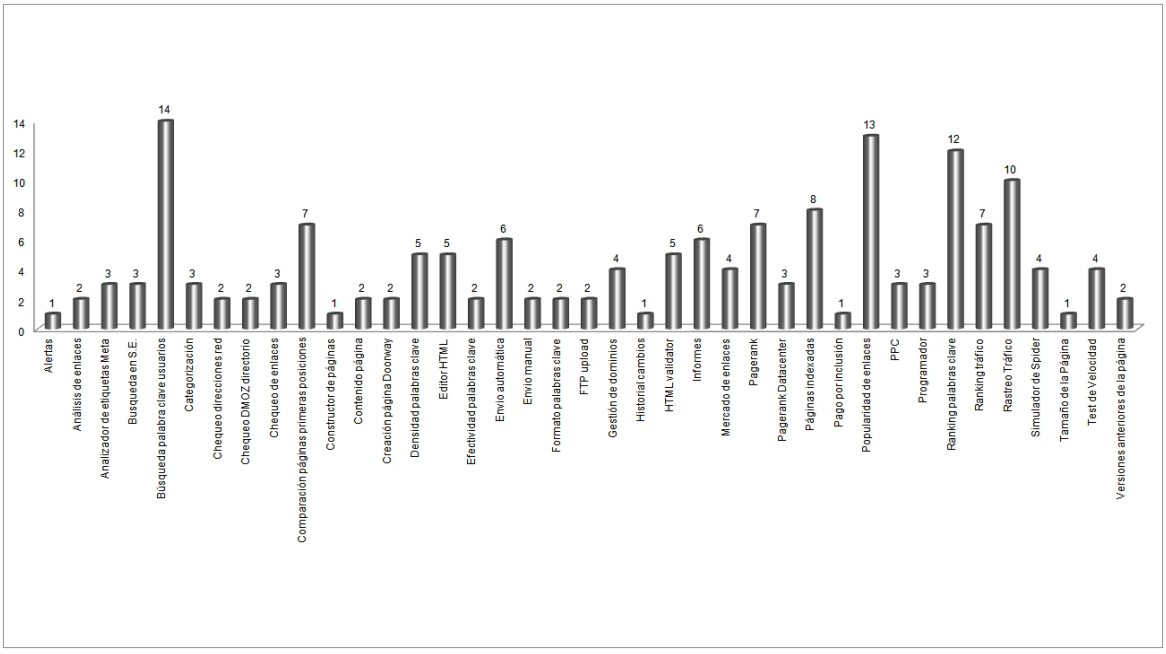
En enero de 2010, para evaluar la capacidad predictiva de las herramientas SEO y la adecuación de los factores a la mejora del posicionamiento, se realizó un estudio con el objetivo de determinar los factores de posicionamiento más discriminantes en relación a las estrategias descritas en la literatura.

1. Técnicas para recolectar factores de posicionamiento

El primer paso consistió en determinar qué factores eran más relevantes para el posicionamiento. En total se identificaron 55 factores, 20 externos a la página (*out the page*) (Tabla II) y el resto presentes en la página web (*in the page*) (Tabla III). La mayoría de los factores se extrajeron de las páginas mediante un desarrollo software propio. Para los elementos externos a la página se recurrió a herramientas SEO y a la información proporcionada por los buscadores. En concreto, las aplicaciones software utilizadas fueron: *SEO Tools*, *URL Trends, Alexa* y *Archive.org*, además de la información suministrada por los buscadores *Google, Yahoo!Search, Bing y Alex*a.

Existen factores de optimización en la literatura que raramente se encuentran en las herramientas SEO. Un ejemplo es la medida en que el término de consulta está presente en los enlaces externos que

Figura 1. Funcionalidades usadas en las herramientas SEO



apuntan al recurso a posicionar, o en qué grado se parece un recurso a otros más antiguos. Por otro lado, como se ha comentado previamente, muchas funcionalidades de las herramientas SEO son ayudas para *webmaster* más que a la optimización, por lo que no están incluidas en las Tablas III y IV. Por último, hay que señalar que los cálculos entre herramientas pueden diferir, como es el caso de recuento de enlaces entrantes entre diferentes herramientas como *Google*, *Url trends* o *Majestic SEO*.

2. Selección de consultas y recolección de datos

Se han seleccionado cinco consultas y se han analizado 200 páginas web en total escogidas entre los resultados. Las páginas analizadas son las situadas entre la posición 1-20 y entre la posición 100-120, para cada consulta. El estudio se centró en recursos que contuvieran lenguajes tipo HTML o XML, descartándose los recibidos por *streaming* o que careciesen de texto. Se ejecutó la búsqueda y de cada recurso se recopilaron los factores de posicionamiento. Por último se almacenó el resultado para una posterior evaluación.

En el proceso de selección de consultas se procuró que, o bien tuvieran un gran número de resultados y variabilidad (consultas genéricas), o bien tuvieran los factores de posicionamiento conocidos (consultas optimizadas). Las consultas realizadas fueron:

- "Documentamania" es el producto de un telendro resultado de las prácticas de posicionamiento en la docencia universitaria de una materia impartida por los autores. El término no estaba presente en la Web antes de plantear la práctica en el curso 2004/05 (sites.google.com/site/documentamaniaproject/). Las páginas web con el término estaban optimizadas con los criterios de posicionamiento conocidos en el momento de su creación.

Tabla II. Factores externos a la página en relación a los algoritmos y las funcionalidades SEO

Factores externos a la página	Descripción	J48	CSE	Fun SEO
Alexa Rank	Cálculo del tráfico que recibe la página web de estudio.		✓	✓
Archive.org_listed	Indica si la página ha sido incluida en *archive.org* (booleano)		✓	✓
Dificultad de posicionamiento	Dificultad de posicionarse con las palabras de búsqueda (*SEO tools*).			✓
Dmoz_listed	Variable booleana que indica si la página de estudio ha sido explorada e incluida por el directorio DMOZ.		✓	✓
Enlaces de la página	Enlaces entrantes en *Google (url trends)*.		✓	✓
Incoming alexa links	Nº de enlaces recibidos contabilizados por *Yahoo*.	✓	✓	✓
Incoming google links	Nº de enlaces recibidos contabilizados por *Google*.	✓	✓	✓
Incoming msn links	Nº de enlaces recibidos contabilizados por *Bing*.			✓
Incoming yahoo links	Nº de enlaces recibidos contabilizados por *Yahoo*.	✓	✓	✓
Listados en *archive.org*	Nº de páginas del site que están incluidas en archive.org.			✓
Nº búsquedas usuarios	Nº de búsquedas registradas para la palabra clave (*SEO tools*).			✓
Nº páginas indexadas	Nº de páginas que indiza *Google* en determinado *site*.			✓
Online since	Nº de días que lleva disponible en la red (*url trends*).			✓
Overall incoming links	Nº total de enlaces entrantes contabilizados en los buscadores		✓	✓
Pagerank	*Pagerank* (*Google)*.		✓	✓
PR por link	Ratio de *Pagerank* de la página entre los enlaces salientes (*url trends*).			✓
Ratio	Ratio que muestra la proporción del número de enlaces de la página web y el *Pagerank* (*url trends*).		✓	✓
Resultados de la palabra clave	Número de resultados devueltos por *Google* a una consulta.			✓
Sitios similares	Páginas relacionadas según *Google*.			
Unique links	Número de enlaces únicos contabilizados por un buscador (*url trends*).	✓	✓	✓

Tabla III. Factores internos a la página en relación a los algoritmos y las funcionalidades SEO

Factores en la página	Descripción	J48	CSE	Fun SEO
Body text	Densidad de la palabra clave en el cuerpo del HTML.			✓
Codificación	Codificación de los caracteres del código HTML			
Content	Tipo de codificación de la página web.			
Densidad alt	Densidad de aparición de la palabra clave en el atributo ALT del HTML.			
Description	Porcentaje de aparición de la palabra clave en la etiqueta *Meta Description*.			
Días última modificación	Fecha en la página de la última modificación.		✓	
Distancia entre palabras	Distancia en el texto de las palabras del término de consulta.			
Enlace externo inaccesible	No es posible conectar con la otra página, se comunica al usuario un error.			✓
Enlace externo no similar	Nº de enlaces externos a sitios con los que no se comparte temática en cuanto a la palabra clave estudiada.			
Enlace externo similar	Nº de enlaces a páginas fuera del dominio que comparten temática mediante las palabras clave.		✓	
Enlace interno	Nº de enlaces que apuntan a páginas del mismo *site*.			
Frame	Presencia de *frames*.		✓	
H1	Palabra claves en las etiquetas de encabezamiento tipo H1.			✓
H2	Palabra claves en las etiquetas de encabezamiento tipo H2.			✓
H3	Palabra claves en las etiquetas de encabezamiento tipo H3.			✓
Keywords	Porcentaje de aparición de la palabra clave en la etiqueta *Meta Keyword*.		✓	✓
Language	Identificador de idioma en el código HTML.			
Meta refresh	Redirección.			
Número img	Nº de imágenes en la página.	✓		
Outgoing links	Nº de enlaces salientes de la página web (*url trends*).			✓
Primera aparición	Distancia en palabras entre el inicio de la página y la primera aparición del término de consulta.		✓	
Ratio código vs texto	Proporción entre el tamaño del código fuente de la página y el texto realmente útil para el usuario (*SEO tools*).	✓		✓
Rel. enlace ext.	Enlaces de relación que apuntan a páginas residentes en otros dominios diferentes al de la página de estudio.			
Rel. enlace int.	Enlaces de relación a páginas del mismo dominio.	✓		
Relación enlace	Número de enlaces de relación (internos al recurso).			
Robots	Si existe la etiqueta meta para los robots o robots.txt.		✓	✓
Style type	Página de estilo que define la presentación de la página.			
Tamaño de la pagina	Tamaño en bytes de la página web (*SEO tools*)	✓		✓
Tamaño del código fuente	Tamaño en bytes del código fuente (*SEO tools*)		✓	✓
Tamaño del Texto	Tamaño de la fuente en la página (*SEO tools*).		✓	✓
Texto	Porcentaje de aparición de la palabra clave en el texto.			✓
Texto enlace	Densidad de aparición de la consulta en texto de enlaces.			✓
Texto párrafos	Densidad de aparición de la palabra clave en los párrafos.	✓		✓
Title	Porcentaje de aparición de la palabra clave en el título.	✓		✓
Type	Lenguaje de scripts de los contenidos del elemento.			

- "Documentamania Universidad". Se precisó más la búsqueda para comprobar cómo afectaba una búsqueda más específica. De nuevo los recursos estaban optimizados.
- "Universidad". Esta consulta genérica proporciona un gran número de resultados. Cada Universidad alberga un elevado volumen de páginas, por lo que es previsible cierto volumen de enlaces internos.
- "Buscador". Es una búsqueda de tipo genérica e imprecisa.
- "Trabajo". Se trata también de un término genérico que debía devolver un gran volumen de resultados dada la amplitud de significados y contextos en los que puede aparecer. Presuponemos que es un concepto potencial a estar sujeto a técnicas de posicionamiento.

Para ejecutar las consultas se seleccionó el buscador *Google*, por acaparar la mayor cuota de mercado en las consultas demandadas como muestran los informes de Hitwise (2008, 2011). *Google* ha mantenido un 65% del mercado en los últimos cinco años, incrementándose hasta un 90% en varios países europeos (Rooney, 2012).

3. Análisis de los datos

Aplicamos minería de datos a la información para crear modelos que expliquen la variabilidad de los factores extraídos en cada consulta. Para el proceso de selección de atributos se aplicó el filtro conocido como *CfsSubsetEval*, junto al método de búsqueda *Bestfirst* (Witten y otros, 2011). *CfsSubsetEval* (CSE) es un algoritmo que evalúa subconjuntos de atributos según la calidad de las predicciones de cada variable. Los subconjuntos de atributos más relacionados con la clase y con menor nivel de correlación con el resto son los mejor valorados. El método de búsqueda *Bestfirst* determina la forma de crear los conjuntos de forma eficiente, incorporando las variables progresivamente. De este modo, el método va analizando la mejora o el deterioro para cada grupo de variables al añadir una nueva variable. Según CSE, los atributos seleccionados como más relevantes fueron: *Incoming Google links, incoming Alexa links, overall incoming links, ratio,* enlaces de la página (*Google*), *keywords*, *pagerank*, *incoming yahoo links*, *unique links*, *DMOZ listed*, tamaño del texto, *title*, *robots*, enlace externo similar, primera aparición, *alexa rank*, *frame*, días última modificación, *incoming yahoo links*, listados en archive.org, tamaño del código fuente. Se crearon clasificadores para observar el comportamiento con cada consulta e identificar las posibles correlaciones e interdependencias que generen modelos. El algoritmo clasificador utilizado es J48. Se trata de una implementación del algoritmo C4.5 propuesto por Quilan (Witten y otros, 2011). Con frecuencia se utiliza en este tipo de estudios por la sencilla interpretación del árbol de aplicación de reglas.

En la tabla IV se muestran los resultados del modelo en cuanto a las instancias correctamente clasificadas. También se muestran las tasas de exhaustividad y precisión, y los factores más influyentes en el modelo. Hay que destacar que la medida de exhaustividad y precisión no hace referencia a si los recursos son relevantes o no a la consulta, sino a la medida en que los factores de posicionamiento ayudan a posicionar páginas en un buscador web. Existe una mejora en el posicionamiento con un incremento de la F de 0,63 a 0,88. Por otro lado, es destacable que en algunos casos las instancias correctas no tienen un modelo subyacente lo suficientemente aceptable, F=0,63, lo cual es claramente

Tabla IV. Resultados de los modelos para evaluar la eficacia de factores de posicionamiento para cada consulta

Consulta en buscador	Instancias correctas	*Recall*	*Precision*	F (media armónica)
Documentamania Universidad	34 (85%)	87%	87%	0,87
Universidad	35 (87,5%)	90%	86%	0,88
Buscador	31 (77,5%)	74%	78%	0,76
Trabajo	28 (70%)	64%	73%	0,68
Documentamania	26 (65%)	60%	67%	0,63

insuficiente para un modelo de clasificación binario. También es relevante que al analizar los factores, estos varían tanto en tasas de éxito como en la naturaleza de los factores de una forma acusada. Se trata de un dato definido en la literatura SEO, donde se establece que cada consulta debe optimizarse de forma particular, sin existir una optimización universal (Moreno y otros, 2011).

El algoritmo J48 crea árboles de reglas que clasifican las instancias (las páginas) de forma dicotómica. Se analiza que factor se tiene en cuenta en cada regla y sí se pueden conocer cuáles son los factores más importantes para clasificar las páginas. Los factores más relevantes según el modelo de reglas con el algoritmo J48 son: *incoming alexa links*, *incoming google links*, *incoming yahoo links*, *rel. enlace int*, *unique links*, *pagerank*, listados en *archive.org*, *title, frame, alexa rank*, número *img*, ratio código vs texto, texto párrafos, tamaño de la página. Reflejan que factores tradicionales como los enlaces entrantes o el término de búsqueda en el título mantienen su importancia.

En las tablas II y III se muestran los factores identificados por los modelos, y también si el factor tiene una funcionalidad equivalente de las identificadas en las herramientas SEO (tabla I). De los 55 factores estudiados el 65,5% (36) fueron identificados como funcionalidades en las herramientas SEO (95% de los externos y el 47% de los internos). Se encontró que 33% de las funcionalidades estudiadas eran relevantes según el experimento (20% de los externos y el 40% de los internos). En la comparación de las herramientas SEO con los atributos relevantes se vio que el 67% era común (100% externos y 57% de los internos). Seis funcionalidades internas (17%), no fueron identificadas en las herramientas SEO, aunque sí estaban presentes como atributos relevantes.

Las diferencias observadas entre factores internos y externos pueden estar relacionadas con la necesidad de los motores de evitar ser manipulados con los factores en la página. Este proceso podría deberse a la asignación de menores pesos a estos factores, como indica la observación de que entre el algoritmo J48 y CSE no haya coincidencias, a diferencia de los factores externos a la página.

Fase 3: Estrategias emergentes de posicionamiento

Continuamente aparecen nuevos comentarios sobre las actualizaciones y diferentes estrategias en los algoritmos de posicionamiento en foros especializados y en los blog oficiales de los buscadores (como *Search Engine Watch*, *seochat*, *blog seomoz* y *blog wordtracker,* y foros de los buscadores como *Google* y *Bing*).

Las principales novedades están dirigidas a potenciar estrategias como la publicidad de algoritmo de posicionamiento, y recursos para mejorar las habilidades del usuario como la personalización de resultados y la integración. También se potencian factores que de algún modo ya se consideraban como la novedad del recurso, el impacto de la web social, la semántica, el prestigio de la fuente y el volumen de información.

Para determinar si las herramientas SEO tienen en cuenta los nuevos factores relevantes en el posicionamiento se examinaron once aplicaciones SEO: *AddWeb Website Promoter*, *Agent Web Ranking*, Herramientas *Google*, *Internet Business Promoter*, *SEO Administrator*, *SEO Elite*, *SEO Tools*, *SEO Open*, *Toolbar Browser*, *Traffic rankings* de *Alexa y Web CEO.* Quedaron fuera del estudio herramientas con funcionalidad limitada como *Google Rankings, Batch HTML Tidy* o *1st Position*. Fueron desestimadas principalmente por restricciones de uso o falta de accesibilidad. Herramientas como *Flash Marketing's Spider, Good Keywords Gold, Web Position Platinium, Search Engine Commando* fueron excluidas por su poca optimización para los factores de análisis. La verificación del acceso instantáneo precisa de un control del buscador. Solo las herramientas que median en la búsqueda pueden acceder a esta información, como es el caso de *Alexa* y *Google Analytics*. Se estudiaron de forma agregada las herramientas de *Google*: *Webmaster tools, Google sitemaps, Google trends, Google analytics, Google suggest, Google instant* y *Rich snippets testing tool*.

Para este estudio hemos analizado seis factores emergentes: 1) la personalización de los resultados, 2) la novedad del recurso, 3) la presencia en la web social, 4) web semántica y vocabularios, 5) el análisis del prestigio del dominio y 6) la valoración del volumen del sitio. No se contempló el factor sobre el procesamiento del lenguaje y el acceso instantáneo, debido a su complejidad, y a su escaso impacto actual en las herramientas SEO.

En cuanto al resultado sobre la personalización de un recurso, se ha observado que algunas herramientas como *SEO Administrator*, *AddWeb Website*, *Google Analytics* o *Alexa* analizan el historial de búsqueda para conocer el comportamiento de un usuario concreto. En el caso *Agent Web Ranking* se incide en funcionalidades relacionadas con el posicionamiento en más de un idioma. Y aunque, la personalización de los resultados es uno de los aspectos más considerados por los buscadores, las herramientas no suelen considerar en la mejora de la posición supuestos como optimizar las búsquedas para determinada ubicación, sistema operativo, dominio, etc.

La novedad ha resultado ser un aspecto escasamente utilizado. *Agent Web Ranking* lo utiliza mediante la comparación entre fechas. *Google Analytics* conserva el historial de visitas, en el caso de SEO tools (módulo *Domain Age Check*) y *Seo Open* se utiliza el repositorio *Wayback Machine*, que almacena un gran número de versiones de sitios web a través del tiempo.

Tabla V. Presencia de nuevas estrategias de posicionamiento en herramientas SEO

Herramientas SEO	Person	Novedad	Acceso instantáneo	Web social	Web Semántica	Prestigio	Volum
Agent Web Ranking	✓	✓					
Herramientas Google	✓	✓	✓	✓	✓	✓	✓
Internet Business Promoter			✓	✓		✓	
SEO Administrator	✓			✓			✓
SEO Elite			✓	✓			
SEO Open		✓		✓		✓	
SEO Tools				✓	✓		
Toolbar Browser		✓		✓	✓		
Traffic rankings Alexa	✓		✓	✓			
Web CEO			✓	✓		✓	

En cuanto al factor relacionado con la web social, *Alexa* y *Url Trends* fueron pioneros, la primera ponía especial interés en las opiniones de los usuarios. También *Google Analytics* podría considerarse en este apartado, al tener en cuenta el origen del tráfico. Este elemento es destacable en *SEO Open* (*Quarkbase*), y en la herramienta *Toolbar Browser* mediante la medición de popularidad. Por su parte, IBP ha ampliado el envío automático a buscadores con el envío a un gran número de redes sociales. En la tabla V aparece muy representado debido a que es un aspecto incorporado tradicionalmente mediante análisis de tráfico o popularidad de enlaces.

El análisis del prestigio del domino es, por el momento, un aspecto tangencial, salvo en herramientas como *Alexa* que de algún modo lo tienen en cuenta. En el caso de *Alexa*, en realidad, se evalúa mediante las visitas de los usuarios. Este aspecto también se muestra en la importancia otorgada por varias herramientas como el obtenido por el directorio DMOZ.

Por último, la valoración del volumen del sitio es un aspecto escasamente incorporado en las herramientas SEO, al menos por el momento. Algunas herramientas como *SEO Administrator* tienen funcionalidades orientadas al análisis de *snippets*. El tamaño de la página también es utilizado por *SEOpen*. Pero quizás sean las herramientas facilitadas a *webmasters* las que más inciden en este aspecto, como *Rich Snippets Testing Tool* de *Google*, dentro de su *Webmaster Tool*.

El caso de las herramientas SEO de Google es particular puesto que la dependencia entre el buscador y sus herramientas es evidente y, sin duda, constituyen una buena opción como estrategias de optimización puesto que incorpora sus tendencias de posicionamiento. Por otro lado, su análisis está condicionado por su propia base de datos como en los aspectos de personalización o novedad de los contenidos. La política de privilegiar ciertos recursos propios en las búsquedas (caso de G+) o de optimizar la búsqueda con el historial de búsqueda de sus usuarios son ejemplos de información inaccesible para herramientas SEO externas.

DISCUSIÓN

El estudio observa la adaptación entre buscadores y herramientas o estrategias de optimización, similar a la competencia entre especies. Evidenciando que una situación estática en los motores web no es realista, ya que aumentaría la vulnerabilidad al *spamming web* por SEO deshonesto y la posibilidad de emulación por otros buscadores de la competencia.

Pese a los reiterados anuncios de los distintos motores por evitar la manipulación deshonesta de las herramientas SEO, las variables sujetas a una mayor manipulación, como las presentes en la página, siguen explicando en parte el posicionamiento logrado. Así, persiste la importancia en el posicionamiento de los enlaces entrantes, la densidad del término de búsqueda en la página, la ratio de código frente al texto o el título. No obstante, sigue siendo incuestionable el mayor peso de otros factores como la edad del recurso, la presencia en sitios de prestigio, el tamaño del recurso y el tráfico.

La satisfacción del usuario parece ser uno de los ejes prioritarios en la evolución de los motores. En este progreso se perciben dos elementos. Por un lado, la preocupación de los motores de búsqueda por el desarrollo de recursos web de calidad, con la información principal fácilmente localizable tanto

para el usuario como para el motor (por ejemplo mediante microformatos consensuados para los *rich snippets*). Por otro lado, el intento continúo de evitar el *spamming web*.

Entre los factores emergentes, hemos encontrado los criterios relacionados con el volumen de información y la novedad. El riesgo de manipulación es obvio, pudiendo realizar por ejemplo una mejora de la novedad mediante modificaciones periódicas de un recurso sin añadir contenido informativo significativo, con el único fin de mejorar la frecuencia de actualización. El resto de factores tienen un mayor grado de independencia del recurso, por estar centrado en la comunidad de usuarios o en fuentes externas. Como ya se ha expuesto, la manipulación malintencionada se dificulta cuando se trata de factores externos y cuando involucra un mayor número de usuarios.

La personalización de resultados es un factor difícil de abordar desde la perspectiva de las herramientas SEO. Aunque facilitar anuncios personalizados tiene un innegable interés desde el punto de vista de los motores o redes como *Facebook*, la hipótesis de que un usuario siempre buscará con los mismos objetivos resulta al menos cuestionable. La implantación de esta estrategia también podría implicar que la información se vaya sesgando con el tiempo, empeorando los resultados hasta presentar incluso dilemas éticos en opinión de algunos autores (Pariser, 2011). No obstante, un estudio de la *London University* mostró que la personalización tiene un impacto limitado, ya que afecta solo a los primeros resultados (Feuz y otros, 2011).

Sobre la implantación definitiva de estos nuevos desarrollos en los motores, anunciados en sus blogs oficiales, hay que subrayar que su implementación en herramientas SEO puede presentar inconvenientes. Muchas de estas novedades están aún en pruebas, por lo que una actualización de las herramientas para incorporar la optimización sugerida por estas, podría carecer finalmente de sentido. Un ejemplo fue el debate suscitado por la ineficacia de herramientas SEO ante *Google Instant* (Allen, 2010) que perdió interés dada su escasa incidencia real. Este hecho fue corroborado posteriormente en el estudio de *eye tracking* realizado con *Rosetta*. Mostraba que, a diferencia de la funcionalidad de autocompletar, los resultados ofrecidos por *Google Instant* eran ignorados en una gran proporción (Miller, 2011).

CONCLUSIONES

Este trabajo ha estudiado los factores de posicionamiento más relevantes considerados por las herramientas SEO. El resultado proporciona una idea de cuáles son factores que constituyen los algoritmos de posicionamiento de los motores de búsqueda web. Las funcionalidades más extendidas son la búsqueda de palabras clave de los usuarios, el ranking de palabras clave y la popularidad de los enlaces. Mediante técnicas de minería de datos se ha analizado la adecuación de las herramientas SEO a la mejora del posicionamiento del recurso.

Hemos comprobado que, en su evolución histórica, el posicionamiento en buscadores web comienza con factores internos que permiten una buena estimación de la adecuación de una página para la consulta. Sin embargo, resultaron ser fácilmente manipulables por los *webmaster*. Un intenso trabajo de actualización continua de los algoritmos de posicionamiento se realiza para combatir la manipulación en las listas de resultados. Sin duda, uno de los hitos para evitar esta manipulación fue usar enlaces externos para calcular la relevancia de una página. En la misma línea se ha observado que los buscadores web tratan de perfeccionar sus algoritmos con nuevos factores externos, y que las herramientas SEO aún no se han acondicionado a los cambios, bien por tratarse de factores emergentes, bien por ser factores difíciles de capturar.

En lo relativo a la evolución constante, los resultados de los buscadores y la optimización racional de los resultados podrían sufrir un deterioro como en el caso del sesgo por información personalizada. No obstante, la mayoría de los cambios en los algoritmos parecen perseguir una mejora de los contenidos y del acceso a los mismos. En este sentido las herramientas deberían evolucionar para capturar, progresivamente, la información que los algoritmos de posicionamiento vayan incorporando.

AGRADECIMIENTOS

Nuestro agradecimiento a Javier Vicent por su aportación y al proyecto HAR2011-27540.

BIBLIOGRAFÍA

Allen, J. (2010). 7 Reasons Why Google Instant Makes SEO Dead-on Relevant. 13 /09/2010. *Search Engine Watch*. Disponible en: http://searchenginewatch.com/article/2050595/7-Reasons-Why-Google-Instant-Makes-SEO-Dead-on-Relevant [20/11 /2011].

Arbildi, I. (2005). Posicionamiento en buscadores: una metodología práctica de optimización de sitios web. *El profesional de la información*, vol. 14 (2), 108-124.

Baeza-Yates, R.; Ribeiro-Neto, B. (2011). *Modern Information Retrieval: the concepts and technology behind search*. (2ª ed.) Harlow (Reino Unido); Addison Wesley, p. 944.

Brin, S.; Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine. *Proceedings 7th International Conf. World Wide Web 7*, p.107-117. Brisbane, Australia: Elservier.

Feuz, M.; Fuller, M.; Stalder, F. (2011). Personal Web searching in the age of semantic capitalism: Dianosing the mechanisms of personalitation. *First Monday*, vol. 16 (2). Disponible en: http://www.uic.edu/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/3344/2766 [7/2/2011].

Flynn, L.J. (1996). Desperately Seeking Surfers. *New York Times*, 11 noviembre 1996. Disponible en: http://www.nytimes.com/1996/11/11/business/desperately-seeking-surfers.html [20/11/2011].

Gyöngyi, Z.; Garcia-Molina, H. (2005). Web spam taxonomy, *Proc. First International Workshop on Adversarial Information Retrieval on the Web (AIRWeb)*. NY: ACM, Disponible en: http://airweb.cse.lehigh.edu/2005/gyongyi.pdf [12/3/2012].

Hitwise, (2008). Google Receives 68 Percent of U.S. Searches in May 2008. Disponible en: http://www.hitwise.com/us/about-us/press-center/press-releases/archived-press-releases/leader-record-growth/ [20/11/2011].

Hitwise, (2011). Bing-powered share of searches reaches 30 percent in March 2011. *Experian Hitwise reports*. Disponible en: http://www.hitwise.com/us/about-us/press-center/press-releases/experian-hitwise-reports-bing-powered-share-of-s/ [11/4/2011].

Karch, M. (2011). Google Bombs Explained: What in the World is a Google Bomb? Disponible en: http://google.about.com/od/socialtoolsfromgoogle/a/googlebombatcl.htm [8/12/2011].

Kleinberg, J. (1998). Authoritative Sources in a Hyperlinked Environment, *J. ACM,* vol. 46 (5), 604-632.

Kobayashi, M.; Takeda, K. (2000). Information retrieval on the web. *ACM Computing Surveys*, vol. 32 (2), 144–173.

Leighton, H.V.; Srivastava, J. (1999). First 20 Precision among World Wide Web Search Services (Search Engines). *JASIS*, vol. 50(10), 870 – 881.

Long, X.; Suel, T. (2003). Optimized query execution in large search engines with global page ordering. *VLDB '03 Proceedings of the 29th international conference on Very large data bases*, p. vol. 29, 129-140. Berlin: Morgan Kaufmann.

Manning, C.D.; Raghavan, P.; Schütze H. (2008). *Introduction to Information Retrieval*. Cambridge University Press. Disponible en: http://nlp.stanford.edu/IR-book/ [12/3/2012].

Marchiori, M. (1997). The quest for correct information on the web: Hyper search engines. In: *Proceedings of the Sixth International WWW Conference,* p. 1225-1235. Reino Unido: Elservier. Disponible en: http://www.w3.org/People/Massimo/papers/WWW6/ [20/11/2011].

Miller, M. (2011). Google searches use autocomplete most, ignore Google Instant (eye tracking study). *Search Engine Watch.* http://searchenginewatch.com/author/2028/miranda-miller [28/11/2011].

Morato, J.; Sánchez-Cuadrado, S.; Valiente, M.C. (2005). Análisis de las estrategias de posicionamiento en relación a la relevancia documental. *El profesional de la información*, vol.18 (1), 21-29.

Moreno, V. (2005). Interacción entre medidas de popularidad en el posicionamiento web. *El profesional de la información*, v. 14 (2), 100-107.

Moreno, V.; Morato, J.; Sanchez-Cuadrado, S. (2011). *Method and system for estimating the position of a resource*. WIPO Patent Application WO/2011/061356

Pariser, E. (2011). Beware Online 'Filter Bubbles'. *TED 2011* (March). Disponible en: http://www.ted.com/talks/eli_pariser_beware_online_filter_bubbles.html [20/11 /2011].

Rooney, B. (2012). Google Searches for Niche Role in Europe, 23/01/2012. *Wall Street Journal*

Singhal, A. (2011). Giving you fresher, more recent search results. 11/03/2011. *The Official Google Blog*. Disponible en: http://googleblog.blogspot.com/2011/11/giving-you-fresher-more-recent-search.html [3/11 /2011].

Thies, D. (2004). Special Report: How To Prosper With The New Google. *SEO Research Labs*. Disponible en: http://www.seoresearchlabs.com/seo-research-labs-google-report.pdf [20/11 /2011].

Wall, A. (2006). *Search Engine History*. Disponible en: http://www.searchenginehistory.com/ [20/11 /2011].

Wall, A. (2011). How Search Engines Work: Search Engine Relevancy Reviewed. http://www.seobook.com/relevancy/ [20/11 /2011].

Wingfield, N. (2011). Microsoft Fires Back at Google, Calls Copying Claims 'Insulting'. *Wall Street Journal*, 3/2/2011. Disponible en: http://online.wsj.com/ [3/2/2011].

Witten, I.H.; Frank, E.; Hall, M.A. (2011). *Data Mining: Practical Machine Learning Tools and Techniques*. (3ª ed.) Burlington (MA); Morgan Kaufmann, p. 629.

ANEXO I

Herramientas SEO evaluadas

La herramienta SEO 1st Position Version 2.5.2.1 consultada en 01/07/2010 actualmente sin funcionamiento. Las demás consultadas en 20/11/2011.

Herramienta SEO	Disponible en URL
1st Position Version 2.5.2.1	http://www.1stposition.net/
AddWeb Website Promoter 8	http://www.cyberspacehq.com
Agente Web Ranking	http://www.agentwebranking.com/english.htm
Batch HTML Tidy	http://www.trellian.com/webtidy/
Flash Marketing's Spider 1.86	http://www.flashmarketing.com/spider/
Good Keywords Gold	http://www.goodkeywords.com/good-keywords/
Google Analytics	http://www.google.com/intl/es/analytics
Google Rankings	http://www.googlerankings.com/
Google Sitemaps	http://www.google.com/sitemap.html
Google Suggest	
Google Trends	http://www.google.es/trends/
Internet Business Promoter 3.0.3	http://www.ibusinesspromoter.com/
Search Engine Commando	http://www.searchenginecommando.com/
SEO Administrator v.3.11	http://www.seoadministrator.com/
SEO Elite	http://www.seoelite.com/
SEO Tools™	http://www.seochat.com/seo-tools/
SEOpen Toolbar	http://seopen.com/
Toolbar browser	http://www.toolbarbrowser.com/
Traffic Rankings Alexa	http://www.alexa.com/
Url Trends	http://www.urltrends.com/
Web CEO Version 6.0	http://www.webceo.com/
Web position platinum 3.5	http://www.wemex.com/