Universidad de Granada

Recuperación de Información

Práctica 1: Búsqueda de información en la web

Miguel Ángel Torres López y Luis Balderas Ruiz

21 de septiembre de 2018

Índice

1. Formatos de codificación

1.1. Codificación de caracteres.

La codificación de caracteres es un método indispensable en la informática moderna, ya que nos permite comunicar y almacenar caracteres desde la representación binaria. No obstante, las técnicas de codificación de caracteres son algo más antigua, un ejemplo de ello es el código Morse.

Desde el punto vista técnico, entendemos por una codificación de caracteres un sistema por el cual representamos caracteres identificándolos uno a uno con otros símbolos normados.

1.2. Formatos de codificación en informática.

En informática, uno de los primeros formatos de codificación que aparece es el ASCII. Este sistema de codificación usa un código de 7 bits, aunque los sistemas actuales están obligados a tomar 8 bits como unidad mínima de almacenamiento. Cada combinación representa un caracter, es decir, tiene 128 caracteres codificados. Este número era suficiente para representar textos en inglés y números, pero pronto quedó expuesta la necesidad de incluir otros caracteres, por ejemplo las caracteres latinos con tildes y otros signos de acentuación.

Para ampliar la gama de caracteres y al mismo tiempo mantener la retrocompatibilidad UNICODE[9], el consorcio para la estandarización de codificación de caracteres, creo el formato UTF-8. Este formato es compatible con ASCII, pero difiere en el tamaño que ocupa un caracter. En UTF-8 los caracteres pueden ocupar 1, 2, 3 o 4 segmentos de 8 bits, siendo el primer segmento el mismo que el de ASCII. Esta nueva codificiación permite representar hasta 1,112,064 caracteres, incluyendo alfabetos occidentales, orientales, símbolos matemáticos e incluso algunos de tipo privado.
En cuanto a rapidez, el formato ASCII está por encima del UTF-8. Aunque a priori vemos que los caracteres tradicionales ocupan igual en los dos formatos, 1 segmento de 8 bits, hay ocasiones en que las posibilidades que nos da UTF-8 entorpezcan la lectura. Por ejemplo, supongamos que tenemos un string 'ABC'. En sendos formatos el string ocuparía 24 bits. Pero acceder al tercer elemento es más costoso en UTF-8, pues tenemos un tiempo adicional al tener que comprobar de que tamaño son los dos primeros caracteres.

Otra codificiación de UNICODE es el UTF-16. Este formato usa 1 o 2 segmentos de 16 bits por caracter, por tanto ya no es compatible con ASCII. No obstante, para lenguajes orientales, como el japonés, la mayoría de los caracteres pueden representarse con 16 bits, lo que aumenta la rapidez de lectura.

Existe tambien el formato UTF-32, con la obligatoriedad de usar 1 segmento de 32 bits por caracter. Al igual que el anterior, no es compatible con ASCII y además es bastante más pesado e ineficiente, pues muchos bits son desaprovechados. Esto lo convierte en un

formato poco usado en la actualidad.

Por tanto, los dos formatos más comunes en la actualidad son UTF-8 y ASCII, aunque poco a poco se está incorporando el formato UTF-16 por compatibilidad con los ordenadores de origen asiático.

1.3. Sistemas operativos y sus formatos

En esta sección haremos un repaso de los principales sistemas operativos y los formatos de codificiación bajo los que trabajan:

- **Windows**. Fue uno de los primeros sistemas en actualizar el formato de codificación. Actualmente utiliza UTF-16, pero sigue sin funcionar de forma óptima con UTF-8.
- **Linux**. Funciona en su mayor parte con UTF-8, aunque dispone de herramientas para cambiar el formato de un archivo. Una de ellas es el comando de termina iconv, que convierte una entrada con un formato a una salida con otro distinto.
- **Mac**. Al igual que Linux trabaja con UTF-8.

1.4. Temporización de búsqueda.

HORA	BUSCADOR	PALABRAS DE BÚSQUEDA
10:34	Google	Formatos de codificación de caracteres
12 minutos	https://www.w3.org/International/articles/definitions-characters/index.es	
6 minutos	https://es.wikipedia.org/wiki/Codificaci%C3%B3n_de_caracteres	
5 minutos	https://en.wikipedia.org/wiki/Comparison_of_Unicode_encodings	
10:57	Google	Unicode
4 minutos	https://unicode.org/	
11:01	Google	Diferencia UTF-8 UTF-16
2 minutos	https://stackoverflow.com/questions/4655250/difference-between-utf-8-and-utf-16	

2. SEO.

SEO (Search Engine Optimization) o posicionamiento en buscadores es uno de los conceptos claves de la industria digital. Tanto el posicionamiento en sí como la profesión que subyace, el SEO se ha convertido en algo que todo empresario, experto en marketing o incluso *influencer* desea entender, manejar o, en el peor de los casos, contratar. En este pequeño informe tratamos de clarificar qué es exactamente el posicionamiento en buscadores desde un punto de vista técnico y por qué es necesario para los afamados *e-business*, así como cuál es el camino para convertirse en un profesional SEO.

2.1. SEO como herramienta de marketing digital

Tal y como [7] menciona, SEO es "el nombre que se le da a la actividad que intenta mejorar los rankings en las búsquedas", es decir, la posición en la que una página web aparece tras realizar una búsqueda con palabras clave (keywords) relacionadas con dicha web. Hay que añadir que, pese a que nuestra web tenga un gran posicionamiento en buscadores, fruto de un trabajo constante que más abajo desarrollamos, los anuncios (que reportan beneficios a Google u otros buscadores) aparecerán siempre antes. Por tanto, el objetivo del SEO es mejorar la posición de una página dentro de las denominadas *búsquedas orgánicas* o *naturales*. Las empresas tienen especial interés en estar presentes de forma conveniente en buscadores y redes dado que recibirán más visitas y, eventualmente, esas visitas se convierten en ventas o beneficios. Por tanto, a pesar de su peso técnico-informático, SEO es una estrategia de marketing en Internet. Como se puede leer en [5], como estrategia de marketing digital, "SEO considera cómo funcionan los buscadores, es decir, los algoritmos de búsqueda subyacentes que dictan el comportamiento de la búsqueda". Gran papel juegan las palabras clave o *keywords* y los enlaces. Además, en los últimos tiempos se están desarrollando nuevos productos entorno al SEO móvil dado que las búsquedas móviles han superado a las mismas realizadas en PCs.

El auge del posicionamiento en buscadores tiene una gran relación con Google. Desde que Larry Page y Sergey Brin idearon *Backrub* y su algoritmo *PageRank*, basado en la cantidad e importancia de los llamados *backlinks o inbound links* (enlaces que recibe una web desde otra o, en otras palabras, la cantidad de páginas que enlazan con el sitio web a través de un vínculo [6]), los distintos propietarios de páginas se percataron de la importancia del posicionamiento y trataron de 'inflar' su PageRank. Ante los intentos de manipular fraudulentamente el PageRank por medio de SPAM, Google desarrolló en 2012 un algoritmo llamado *Google Penguin*, que evalúa la calidad de los links de los que proceden las páginas web.

2.2. Herramientas SEO

Como dice [8], existen multitud de herramientas que nos ayudan a conocer el posicionamiento en buscadores de una página o negocio online. Entre ellas destacan:

- Google Analytics: Genera estadística con multitud de información interesante para nuestro negocio, como cuándo y cuántos usuarios navegan por nuestra web.
- Webmaster Tools
- Bing Webmaster Tools
- Sitemap XMl Generator
- Herramienta de palabras clave Adwords: Nos ayuda a crear contenido que nos permitan generar textos de calidad y originales.
- Google Trends: Nos ayuda a buscar tendencias entre los usuarios en torno a un producto o servicio

2.3. Técnicas SEO

Las técnicas del posicionamiento pueden ser muy variadas. Según [12], las técnicas SEO se pueden clasificar en dos categorías: White hat SEO: Técnicas recomendadas por los buscadores como buen diseño; y Black hat SEO: Conocidas como spamdexing, no son aprobadas por los buscadores.

2.3.1. White Hat SEO

White Hat SEO son aquellas técnicas éticamente correctas que cumplen las directrices marcadas por los motores de búsqueda para posicionar una web [10]. Más concretamente:

- Sigue las indicaciones del buscador.
- No es engañosa.
- Asegura que el contenido que el buscador indexa es el mismo que el usuario verá finalmente.
- Asegura que el contenido de la web se ha creado por algún usuario.
- Asegura un contenido de calidad.
- Asegura disponibilidad del contenido web.

2.4. Black hat SEO o Spamdexing

El nombre es una contracción del inglés que se refiere a la indexación de spam, como vemos a continuación:

- Sirve una versión de la página para bots y otra para usuarios humanos (Cloaking).
- Repite keywords en las metaetiquetas, así como usa keywords no relacionadas con el contenido de la web (Metatag Stuffing)

- Páginas web de contenido de baja calidad pero manipuladas con palabras clave (Gateway Pages).
- Crear una página web con contenido similar a otra página, pero que redirige a webs maliciosas (Page hijacking)

Hoy en día los buscadores son capaces de identificar todas estas características maliciosas de las indexaciones, por lo que se recomienda no incluir ningún enlace de estas características para mejorar la posición.

2.5. SEO como profesión

Como hemos comentado anteriormente, el posicionamiento en buscadores es una pieza clave en el marketing online de las empresas. Tanto es así que ha dado lugar a una nueva profesión: el especialista SEO. En palabras de [11], los especialistas en posicionamiento no son normalmente programadores web, si no que son promotores de diseños web de calidad. Un SEO debe poseer un gran número de habilidades. Entre ellas se encuentra la capacidad de analizar páginas web, dado que será el responsable de mejorarla para llegar a más gente y de forma más atractiva. De la misma forma, deberá optimizar y actualizar las palabras clave acorde a la tendencia de los usuarios en sus consultas. Como consecuencia de las dos anteriores, el SEO generará el contenido de la web con el objetivo de mantener una web de calidad con información en todos los formatos (textual, gráficos, multimedia...). Los conocimientos de marketing, negocios y habilidades IT (manejo de lenguajes de programación como HTML, CSS u otros) son cruciales ya que se trata de una profesión multidisciplinar, donde se aúnan los intereses económicos con las nuevas tecnologías y la publicidad.
Como hemos comentado anteriormente, el posicionamiento en buscadores es una pieza clave en el marketing online de las empresas. Tanto es así que ha dado lugar a una nueva profesión: el especialista SEO. En palabras de [11], los especialistas en posicionamiento no son normalmente programadores web, si no que son promotores de diseños web de calidad. Un SEO debe poseer un gran número de habilidades, como las siguientes:
Hay ciertos aspectos en los que un especialista SEO se puede centrar. Algunos de ellos son:

- Escritor del contenido. Una ocupación menos técnica y más de marketing,pero crítica para un buen posicionamiento.
- Link builder, de gran importancia ya que elegir links de páginas web reputadas puede ser complicado.
- Investigador de la web, ya que la web es un sistema de información complejo y en constante actualización y crecimiento.

Al ser una profesión tan novedosa no existen titulaciones oficiales, aunque sí multitud de congresos, exposiciones, webinarios y blogs donde formarse. No obstante, existen algunos portales de internet que expiden certificación acreditando la maestría del alumno, como

ExpertRating.com o *SEOPros.org*. Además, los salarios varían dependiendo del cargo, pudiendo llegar a los 100000$ al año.

2.6. Temporización de búsqueda.

SESIÓN DE BÚSQUEDA: SEO

Búsquedas individuales: 6
Minutos: 35

HORA	BUSCADOR	PALABRAS DE BÚSQUEDA
16:30	GOOGLE	SEO
2 mins https://es.wikipedia.org/wiki/Posicionamiento_en_buscadores		
16:32	GOOGLE	SEO english wiki
5 mins https://en.wikipedia.org/wiki/Search_engine_optimization		
16:37	GOOGLE	seo positioning
4 mins - http://www.seodesignsolutions.com/blog/seo/positioning-seo/ - http://www.seodesignsolutions.com/blog/search-engine-optimization/how-to-build-links-and-optimize-your-website-for-multiple-keywords/		
16:41	GOOGLE	what is a SEO
12 mins - https://www.redevolution.com/what-is-seo - https://www.seo.com/blog/what-is-an-seo-specialist/		
16:53	ECOSIA	how to be SEO specialist
7 mins - https://www.marketingcareeredu.org/search-engine-optimization-specialist/		
17:00	ECOSIA	SEO tecnics
5 mins - https://www.tutorialspoint.com/seo/seo-tactics-methods.htm		

3. Detección de plagio

3.1. Definición.

Según *plagiarism.org*[3], se considera plagio la acción de presentar o usar trabajo de otro autor sin citarlo y sin especificar la fuente de procedencia. Según la RAE[4], plagiar se define como la acción de copiar obras ajenas y darlas a conocer como propias.

En el área que nos compete, hablaremos de plagio en texto escrito, aunque tambien se considera plagio el uso de cualquier producción de un autor sin su debida mención. Cabe mencionar el caso de los archivos multimedia. Publicar imágenes, vídeos o fragmentos y composiciones de los mismos de otro autor sin su correspondiente fuente de procedencia también es motivo de conflicto.

3.2. Casos de plagio en la actualidad.

En 2015 se abrió una página de Facebook llamada *Cabronazi* que empezó a publicar fotografías con mensajes graciosos. Tres años después cuenta con más de 12 millones de seguidores y factura cerca de 370000 euros al año. No obstante, detrás de este éxito muchos usuarios se quejan de que la mayoría de las publicaciones son plagio de otras en las redes sociales. Puede verse la discusión en el periódico *El Confidencial*[1] .

Un tema que genera más controversia en la actualidad es el de Pedro Sánchez, al que acusan de haber plagiado ciertas partes de su tesis doctoral. Tras haber publicado en internet dicha tesis, distintos medios han examinado el documento. Estos son los resultados según uno de esos medios[2].

3.3. Métodos para detectar plagio.

Existen numerosos métodos para detectar plagio. Los más frecuentes por su rapidez son los software de detección de plagio, aunque para usarlos se necesita los documentos en formato digital.

- **Cadenas de texto.** La mayoría del software comercial se basa en la comparación de cadenas de texto. Usan una base de datos de documentos para enfrentar el documento sospechoso. Esto plantea un inconveniente, puede que la base de datos usada no sea suficientemente extensa o no contenga el documento plagiado. Se podría intentar añadir la mayor cantidad de textos posibles para mejorar el contraste, pero esto supondría una penalización en tiempo de cómputo.

- **Bolsas de palabras.** Una forma de reducir el tiempo sería usar comparación de bolsas de palabras. Una bolsa de palabras es un vector que representa las palabras de un texto. Por tanto, al usar bolsas de palabras estamos comprobando si dos

textos usan el mismo vocabulario, obviando el orden en el que las palabras aparecen. Esto produce una reducción de la eficacia del método.

Notar que sendos métodos pueden ser mejorados con el uso de diccionarios de sinónimos y traductores para evitar la reformulación de oraciones y las traducciones.

- **Analizador de estilo.** Existen otras técnicas que proveen un análisis más profundo del texto, por ejemplo los analizadores de estilo. Este tipo de detectores analizan distintos aspectos en el discurso de un autor para hacer un perfil de escritura. Este perfil se puede caracterizar por la longitud de las oraciones, el uso de muletillas o de reformuladores del discurso. Cuando se introduce un nuevo documento en el sistema, se realizar un análisis para encontrar estilos similares. Hay que notar que un analizador de estilo es sensible a idiomas, por tanto, las traducciones no literales de textos no serían detectadas por este método.

3.4. Temporización de búsqueda.

HORA	BUSCADOR	PALABRAS DE BÚSQUEDA
18:30	Google	Detección de plagio
2 minutos	https://wwwhatsnew.com/2017/10/07/4-sitios-web-para-detectar-plagios/	
18:32	Google	Plagiarism detection
11 minutos	https://en.wikipedia.org/wiki/Plagiarism_detection	
18:43	Google	Plagiarism
8 minutos	https://www.plagiarism.org	
18:51	Google	Bag of words
4 minutos	https://medium.com/greyatom/an-introduction-to-bag-of-words-in-nlp-ac967d43b428	
18:55	Google	Estilometria
2 minutos	https://www.estilometria.com/	
18:57	Google	Rae
1 minuto	http://dle.rae.es/?id=TIZy4Xb	
18:58	Google	El cabronazi plagio
6 minutos	https://www.elconfidencial.com/tecnologia/2018-08-09/cabronazi-cabroworld-carlos-soria-bernardo-memes_1602201	
19:04	Google	Plagio Sanchez
4 minutos	https://elpais.com/politica/2018/09/14/actualidad/1536938921_232616.html	

Referencias

[1] Artículo sobre el plagio de cabronazi. 2018. (https://www.elconfidencial.com/tecnologia/2018-08-09/cabronazi-cabroworld-carlos-soria-bernardo-memes_1602201/).

[2] Artículo sobre la tesis de pedro sánchez. 2018. (https://elpais.com/politica/2018/09/14/actualidad/1536938921_232616.html).

[3] Página web de plagiarism. 2018. (https://www.plagiarism.org).

[4] Sitio web de la real academia española. 2018. (http://dle.rae.es/?id=TIZy4Xb).

[5] https://en.wikipedia.org/wiki/Search_engine_optimization, consultado el 18 de septiembre de 2018.

[6] https://es.wikipedia.org/wiki/Backlink, consultado el 18 de septiembre de 2018.

[7] https://www.redevolution.com/what-is-seo, consultado el 18 de septiembre de 2018.

[8] https://www.seoalcuadrado.es/seo-que-es/, consultado el 18 de septiembre de 2018.

[9] https://unicode.org/, consultado el 19 de septiembre de 2018.

[10] https://www.40defiebre.com/que-es/white-hat-seo/, consultado el 19 de septiembre de 2018.

[11] https://www.marketingcareeredu.org/search-engine-optimization-specialist/, consultado el 19 de septiembre de 2018.

[12] https://www.tutorialspoint.com/seo/seo-tactics-methods.htm, consultado el 19 de septiembre de 2018.