Fundamentos de Búsqueda y Evaluación de Información para Periodistas y Comunicadores

Por **Lluís Codina, Rafael Pedraza**
Área de Formatos Digitales e Interactivos
Grupo DigiDoc
Departamento y Facultad de Comunicación
Universitat Pompeu Fabra
Enero 2015

1. Buscadores generalistas

Antes de introducir el uso de los buscadores vale la pena considerar cómo funcionan y cuál es la anatomía de un buscador. El objetivo es tener un grado de comprensión de su naturaleza, aunque sea a mil metros de altura, que nos ayude a utilizarlos mejor. En realidad es difícil imaginar el papel de los medios actualmente sin su relación con los buscadores. Conocerlos mejor nos ayudará no solamente a ser más eficientes encontrando información para nuestro trabajo como periodistas o comunicadores, sino también a que nuestras producciones periodísticas tengan mayor visibilidad en la web.

La cuestión es que los motores de búsqueda son los más importantes sistemas de información de nuestro tiempo. Forman parte del paisaje de Internet de una manera que parece definitiva. Nadie imagina hoy la posibilidad de explotar la inmensa riqueza de la Web sin el auxilio de los motores de búsqueda cuyas funciones no han dejado ni de aumentar ni de experimentar cambios desde su temprana aparición en los años 90.

Sistemas de información documental

Una forma que puede resultar útil para concebir a los buscadores consiste en considerarlos como una clase de sistema de información documental; los cuales a su vez se pueden representar con el diagrama (Ilustración 1) que mostramos a continuación:

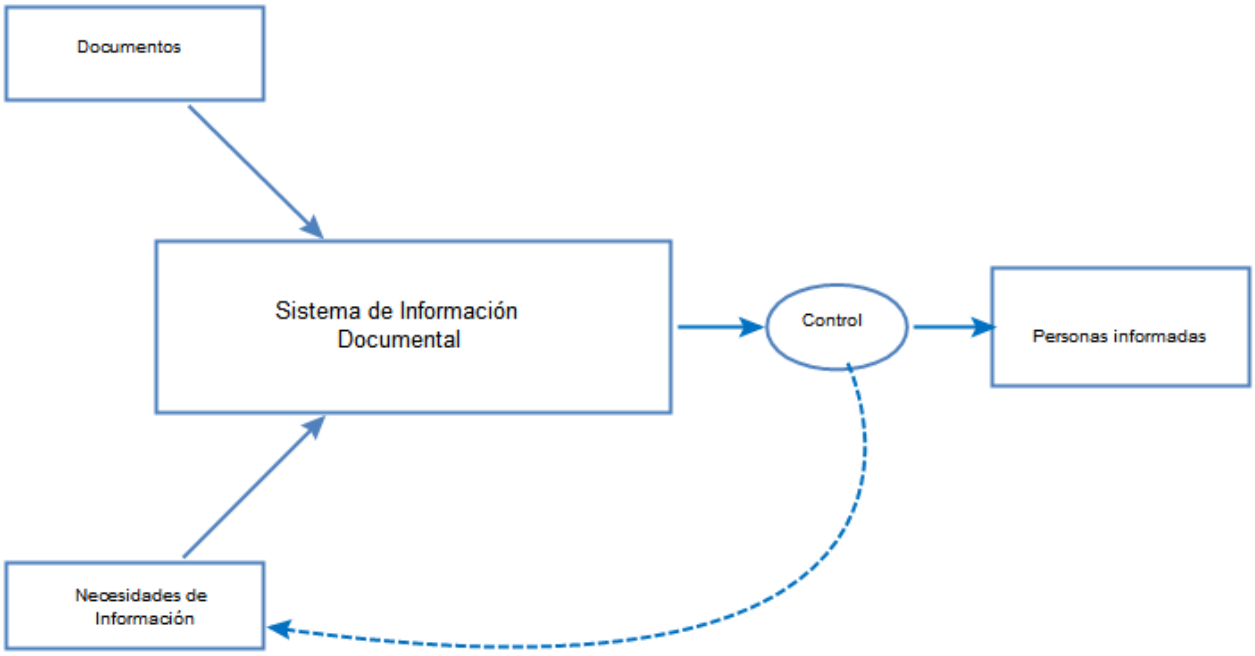


Ilustración 1: Diagrama general de un Sistema de Información documental

Podemos ver que un sistema documental (tal como un motor de búsqueda) acepta dos clases de entradas: Documentos y Necesidades de Información. Los documentos en nuestro caso son páginas web (pero no únicamente, ya que los servidores web incluyen toda clase de documentos) y las necesidades de información son las preguntas de los internautas.

La salida del sistema es una lista de enlaces a sitios web que, presumiblemente, contienen información capaz de satisfacer las necesidades de información del usuario, de aquí que la salida se rotule con el (optimista) texto de Personas informadas. Naturalmente, si la primera consulta no da buenos resultados, lo lógico es que el usuario (el internauta) haga algún intento más. Esto queda representado por la elipse rotulada con la palabra Control. Es decir, se supone que el usuario examina la relevancia de los resultados y si no le satisfacen, reitera la pregunta usando otros términos o parametrizando la consulta mediante la búsqueda avanzada.

Estructura de un motor de búsqueda

Un motor de búsqueda (MdB a partir de ahora) consiste en un conjunto de programas que trabajan en cooperación. Estos programas se pueden distribuir en dos grupos principales:

- *Sistema de exploración,* también llamado *Rastreador* (en inglés: *spyder* o *crawler*)
- *Sistema de recuperación de información* (en inglés: *information retrieval system*)

Cada uno de estos sistemas se compone de otros subsistemas (Ilustración 2):

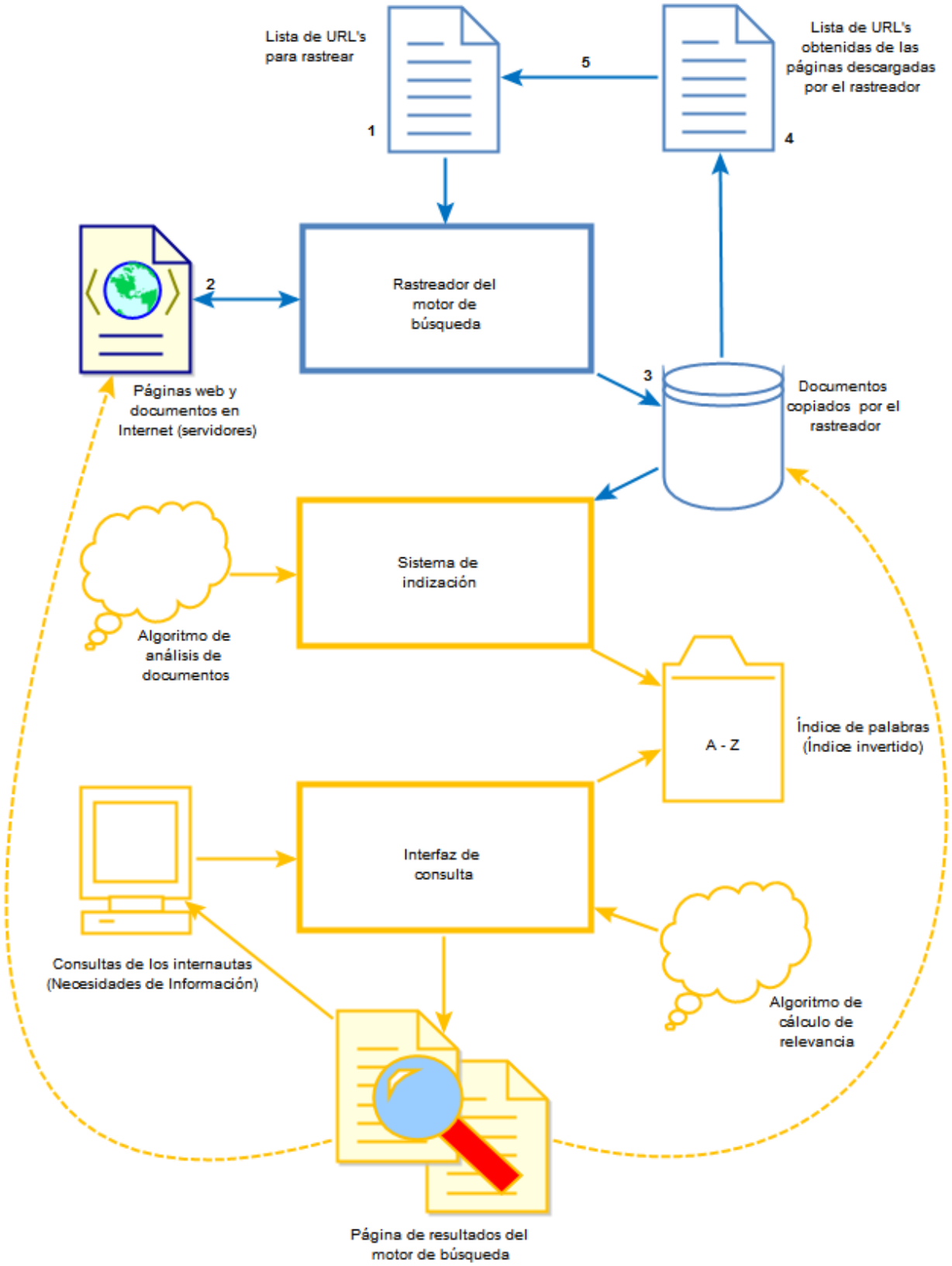


Ilustración 2: Estructura y funciones de un Motor de Búsqueda

Rastreador del motor de búsqueda

El rastreador, o *sistema de exploración* tiene la misión de descubrir y descargar páginas web y otros documentos de Internet. Para ello, parte de una lista inicial de URL o direcciones de páginas web (1). A partir de ella accede a documentos (2) publicados en la web. Descarga esos documentos, en su mayoría páginas web, pero también documentos ofimáticos, presentaciones, archivos de imágenes, etc. en el almacén o repositorio de documentos del MdB (3). A partir de las páginas descargadas, el rastreador obtiene nuevas URL (4) que añade a su lista de URL pendientes de analizar (5). Cuando producimos una página web nueva, nuestra esperanza es que los MdB la detecten y la indicen. Pero no tenemos porqué esperar a que esto suceda. Podemos tomar la iniciativa y avisar a los MdB, y en particular a Google de que tal página existe y conseguir así que la índice.

Actividad 1

Acuda a la versión inglesa de *Wikipedia* (http://en.wikipedia.org/wiki/Web_crawler) y haga una revisión del artículo, aunque sea en diagonal, prestando atención a los diagramas y a la lista de *crawlers* que presenta el artículo. *Alternativa*: Vaya a la página de la Wikipedia dedicada a los rastreadores o crawlers (http://es.wikipedia.org/wiki/Web_crawler) y lea el artículo dedicado a este tema. Haga capturas de pantalla comentadas de la actividad.

Sistema de recuperación de información

El sistema de recuperación de información tiene a su vez dos componentes:

- Sistema de indización
- Sistema de consulta

El primero de ellos se ocupa del análisis de los documentos descargados de la web y de la creación de los índices que permiten luego las consultas al motor; mientras que el segundo constituye la interfaz visible del MDB, precisamente la parte con la que interactúan los usuarios.

Sistema de indización

Si el MDB puede responder preguntas en el asombrosamente breve tiempo al que nos ha acostumbrado tanto Google como Yahoo (típicamente, fracciones de segundo), es porque los motores no exploran la web para los usuarios en tiempo real, es decir, en el momento de la consulta.

Un sistema que pretendiera explorar de forma secuencial los documentos de la web para ofrecer respuestas no tendría el más mínimo sentido. Incluso con la potencia de cálculo de los ordenadores actuales, podrían transcurrir meses entre la pregunta y la respuesta. Un estudiante que hiciera su pregunta en el primer trimestre de la carrera, cuando obtuviese la respuesta ya estaría en el segundo curso (eso si no suspendió el primer trimestre esperando la respuesta del motor). En lugar de lo anterior, lo que hacen los motores es consultar sus índices internos.

Actividad 2: Haga diversas consultas (dos o tres es suficiente) con palabras relativamente habituales (como televisión, ordenadores, música, etc.) en Google o en Yahoo y compare los tiempos de respuestas y el número aproximado de documentos encontrados (ambas cosas indicadas en una barra situada justo sobre la lista de resultados). Luego haga dos o tres preguntas con términos más especializados (como "conductas adictivas", "derecho al olvido", "crisis del euro", etc.) Compruebe si, con diferentes preguntas, los tiempos de respuesta son muy distintos o, en general, los tiempos son similares. Cabe esperar que no haya diferencias apreciables. El motivo se discutirá más adelante. Haga capturas de pantalla comentadas.

Índices directos

Los MdB utilizan al menos dos clases de índices, denominados respectivamente índices directos e índices invertidos. Las consultas las resuelven los *índices invertidos*, mientras que los *índices directos*, que son los que vamos a ver ahora, se utilizan como elementos de gestión y control internos.

En un índice directo tenemos la lista de documentos (o de registros) en un orden cronológico (el documento más antiguo primero, por ejemplo) o numérico (del documento número 1 hasta el último). La siguiente figura ilustra esta clase de índices:

Núm. Documento	*Contenido*
00001	**Título**: Manual de lenguajes documentales **Autor:** Blanca Gil
00002	**Título**: Diseño y construcción de sitios con Dreamweaver 8 **Autor:** Marc Campbell
...	...
12345	**Título**: Aprender a razonar **Autor:** Fina Pizarro

Representación de un índice directo

Como podemos ver, con este índice, para saber si hay un documento con las palabras "aprender" y "razonar", tendríamos que recorrer decenas de miles de entradas del índice (concretamente 12345). Lo peor es que, si el índice completo tuviera, por ejemplo, cien mil entradas, habría que recorrer las cien mil entradas del índice para saber si hay más de un documento que cumpla la condición anterior.

Como es fácil suponer, esta clase de índices no mejora mucho en relación al supuesto rastreo

en tiempo real de la Web con el que especulábamos antes. Por tanto, un MdB necesita complementar este índice con un índice invertido, que es el que se utiliza realmente para responder a las consultas (mientras que el índice directo se utiliza para aspectos de gestión y administración internos).

Índices invertidos

La estructura de un índice invertido es exactamente la inversa de la anterior (de aquí el nombre), a saber: consiste en una lista ordenada de todas y cada una de las palabras que aparecen en los distintos documentos asociadas a los documentos concretos en los que aparecen. La estructura típica de un índice invertido, por tanto es la siguiente:

Término único	*Frecuencia*	*Ubicación*
Aprender	110	(12345, 01, 01) ...
...	...	...
Blanca	233	(00001, 02, 01) ...
...	...	...
Campbell	6	(00002, 02, 02) ...
...	...	...
Razonar	214	(12345, 01, 03) ...
...	...	...

Representación de un fichero invertido

La explicación:

- En ***Término único*** aparecen las distintas palabras de los documentos, pero solamente aparece una fila por cada palabra (aunque en el conjunto de los documentos aparezca muchas veces).
- En la columna ***Frecuencia*** tenemos el número total de veces que aparece cada término.
- Por último, en ***Ubicación*** tenemos una clave en forma de vector donde aparece el número de documento, la zona o campo donde aparece la palabra y el orden de la palabra. Habrá un vector por cada ocurrencia.

Veamos esto último con más detalle. Tomemos el término “Aprender”. Tenemos que la columna de ***Frecuencia*** señala: “110”**.** Esto significa que “Aprender” aparece 110 veces en el conjunto de los documentos, y que habrá por tanto 110 vectores distintos en la columna ***Ubicación***. Nosotros solamente hemos representado uno de ellos para simplificar. Tomemos ahora este mismo vector [(34567, 01, 01)] y veámoslo con detalle. El primer número, 34567, es el que identifica el documento, el segundo, 01, identifica el campo o zona del documento donde aparece la palabra, en este caso, Título.

Otros índices auxiliares ayudan al MdB a recorrer la lista de términos únicos con un pequeño número de operaciones de comparación. Lo mejor de todo es que esta clase de índices hace que el tiempo de respuesta sea virtualmente independiente del número de términos que aparecen en la lista.

El problema principal con estos índices es la cantidad de recursos de hardware que requieren, particularmente, si se debe dar servicio a miles de usuarios simultáneamente.

Actividad 3

Google utiliza alguna versión del tipo de índice invertido que hemos mostrado. Para ver que estos índices son eficaces, pero no permiten hacer milagros, haga una consulta en Google por un término muy frecuente como hizo antes (p.e. televisión) y compruebe si le resulta posible llegar al resultado número 1.000 (es decir, compruebe si puede llegar a la página de resultados número 100 ya que cada página contiene 10 resultados, 100 x 10=1.000). Probablemente, no podrá. En ocasiones no es posible pasar del resultado 400 o 500. El motivo no está documentado oficialmente, pero es fácil deducir que la red de ordenadores de Google tiene una cantidad de memoria limitada que puede asignar a cada usuario y por tanto no puede cargar toda la información de los resultados. Haga capturas de pantalla comentadas.

Para que el MdB pueda construir su índice necesita a su vez un procedimiento que le permita analizar las páginas web y extraer los términos de las mismas que formarán parte del índice, y para ello a su vez necesita poder determinar qué es y qué no es un término.

El algoritmo más fácil determinaría que son términos todas las cadenas de caracteres situadas entre espacios en blanco. Sin embargo este algoritmo produciría muchos errores. Por ejemplo, no seleccionaría las palabras iniciales de cada frase ni sabría tratar las que estuvieran rodeadas de signos ortográficos; por otro lado, tampoco conviene identificar como términos las palabras sin significado propio, como los pronombres y los artículos, etc.

Interfaz de consulta

Las funciones que desempeña son las siguientes:

- En primer lugar, hay un formulario de consulta sirve para enviar la pregunta a la parte del sistema que compara los términos de la pregunta con el índice invertido para filtrar las páginas web que contienen los términos de la pregunta.

- En segundo lugar, el MdB debe presentar los resultados de la forma más clara y eficiente posible a través de la *Página de resultados del motor de búsqueda (*en las siglas inglesas*,* SERP, por *search engine results page).*

- Por último, pero posiblemente lo más importante de todo, los resultados deben mostrarse en algún orden significativo, y de entrada podemos descartar el orden alfabético o el cronológico, dada su escasa utilidad en un contexto tan heterogéneo y con la inmensa cantidad de documentos que hay en Internet. Aquí interviene el denominado algoritmo de cálculo de relevancia, al que dedicaremos el resto de este apartado.

Actividad 4

Vaya a esta dirección: http://www.google.com/insidesearch/howsearchworks/. Revise, aunque sea en diagonal, los dos primeros apartados, (1) **Rastreo e Indexación** y (2) **Algoritmos**, de este sitio de Google que explica de forma algunos aspectos de su funcionamiento interno. Haga **capturas de pantalla** comentadas de cada uno de los apartados. Vaya después al tercer apartado (3) **Spam**. Revise con mayor atención este apartado. Observe en particular los ejemplos de spam (también llamado blackhat SEO) detectados por Google en tiempo real. Haga capturas de pantalla comentadas.

Cálculo de relevancia

En recuperación de información se considera que los distintos documentos de un fondo presentan un grado de relevancia diferente para cada pregunta o necesidad de información. Ante una pregunta dada, en un extremo tendremos a los documentos cuya relevancia será cero (por no tener ninguna relación con la pregunta). En el otro extremo tendremos (con suerte) documentos cuya relevancia será total (o casi total). En zonas intermedias tendremos documentos con diversos grados de relevancia.

Si la expresamos con tanto por ciento, tendremos un espectro que puede ir del 0% de relevancia al 100% de relevancia con zonas intermedias donde habrá documentos con el 10%, el 25% o el 70% de relevancia, etc.

Lo más lógico será presentar la página de resultados con los documentos ordenados según su grado de relevancia. Esta ordenación se denomina también *ranking.* De este modo, aunque haya miles y miles de documentos relevantes obtenidos como respuesta, podremos limitarnos a consultar los 10 o 20 primeros, siempre que sean los más relevantes. Otros modos de ordenación, como el alfabético o el cronológico no nos permitirían tal economía de consulta. Para estar seguros que accedemos a los más relevantes ¡tendríamos que examinar la lista completa aunque estuviera formada por millones de documentos!

Los MdB actuales suelen combinar dos grupos de criterios para determinar la relevancia de una página web:

- Criterios internos o intrínsecos (*on page*)
- Criterios externos o de popularidad (*off page*)

Criterios intrínsecos – On page

Los criterios internos se refieren principalmente al contenido de la página y muy concretamente a aspectos estadísticos o de frecuencia de ocurrencia de la palabra (o palabras) clave de la pregunta. En concreto, las páginas web que tengan mayor densidad de la palabra clave, serán más relevantes.

Otros aspectos, como que la palabra clave aparezca rodeada de etiquetas como <h1> o <h2> también otorgan mayor importancia relativa a la página correspondiente, y sobre todo, al hecho de que la palabra clave forme parte de la URL de la página. Algunos motores, como Google, indican que utilizan más de 200 indicadores como los mencionados. Otros indicadores incluyen la antigüedad del dominio, el volumen de información, la frecuencia de actualización, el uso de componentes multimedia, etc.

Criterios externos – Off Page

Los criterios externos se refieren, principalmente, al resultado que pueda arrojar el análisis de enlaces del sitio. El análisis de enlace incluye tanto los enlaces de salida (los enlaces de la página considerada hacia otras páginas) como los enlaces de entrada (los enlaces hacia la página considerada procedentes de otras páginas). De todas formas, al parecer, el peso mayor recae

en los enlaces de entrada. En este sentido, en general, una página es más importante cuántos más enlaces de entrada tiene procedentes de otras páginas.

Sin embargo, no todos los enlaces valen igual. Los enlaces de entrada de páginas que, a su vez, son muy enlazadas, son más importes. Es decir, si la página B está enlazada por otras 50, y en cambio la página A está enlazada por otras 75, la página A es más importante que la B a igualdad de importancia de las páginas que envían el enlace. En cambio, si los enlaces que recibe la página B son de páginas, a su vez muy enlazadas, la página B, con menor número de enlaces puede ser considera más importante que la A.

Cada MdB tiene sus propios conjuntos de criterios y sus propias reglas para asignar pesos a cada criterio, pero en general, el número de enlaces que recibe una página y la antigüedad de la misma (un valor positivo) suele ser los más importantes, al menos a igualdad de los otros factores.

Página de resultados - SERP

Por último, la respuesta del motor con la lista de documentos organizados en un ranking debe presentarse de alguna forma. De este aspecto se ocupa precisamente la página de resultados o SERP (por Search Engine Results Page).

Las funciones de la SERP, más allá de poder hacer clic en ellos y, por alguna razón, suelen pasar desapercibidas. Tal vez está bien que sea así, porque cada vez que se han hecho intentos en contra, parece el motor ha perdido popularidad.

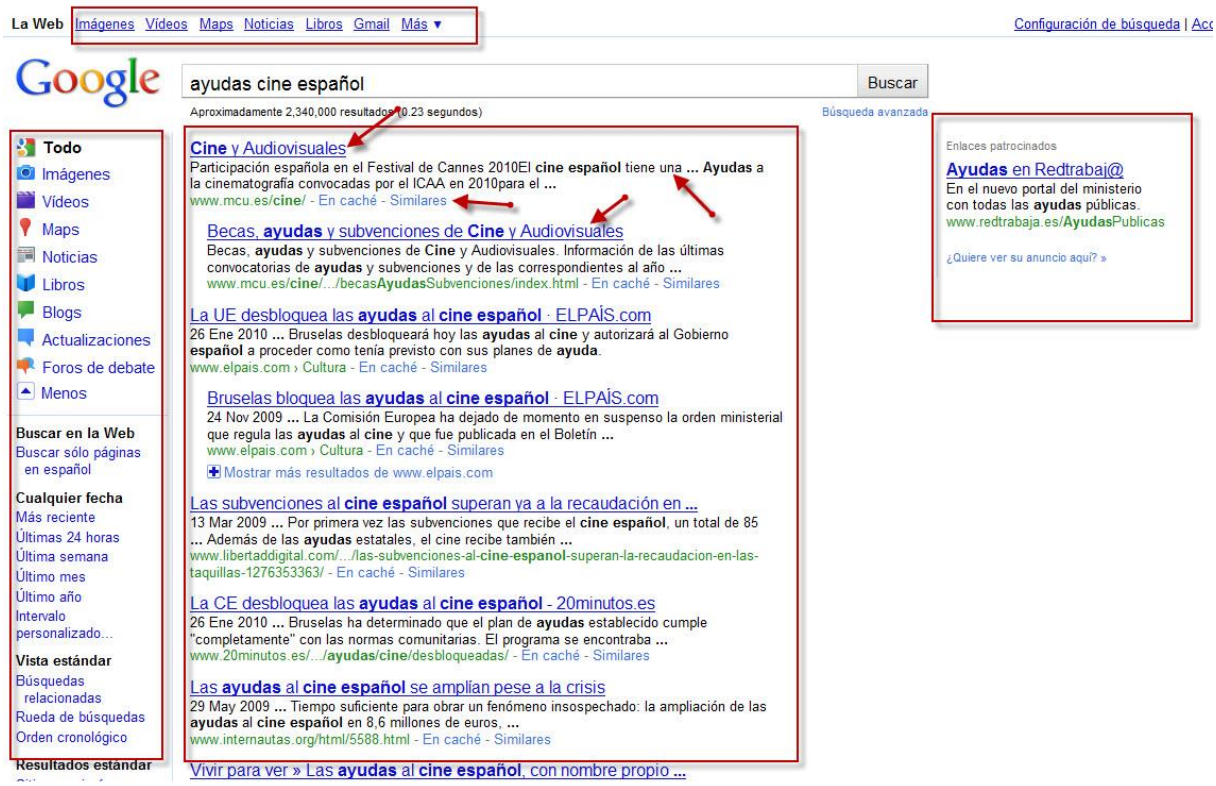
En este sentido, a simple vista, la SERP ofrece una lista de páginas o documentos y ya está. Sin embargo, una mirada atenta a la SERP de, por ejemplo, Google, nos muestra las siguientes informaciones y opciones:

- Número total de páginas que contienen la palabra clave.
- La sección de resultados de pago (resultados patrocinados o más claramente publicidad). De estos resultados vemos un pequeño anuncio, con un título. Una frase descriptiva y una URL.
- La sección de resultados no pagados (resultados orgánicos). De cada uno de estos resultados, tenemos la siguiente información:
    - Título de la página (o del documento)
    - El tipo del documento (cuando no es HTML)
    - Unas líneas de descripción del contenido del documento
    - URL de la página
    - Tamaño de la página web (no de los archivos enlazados desde ella)

- Y las siguientes opciones:
    - Obtener una versión traducida de la página con traducción automática (o sea, en general muy deficiente)
    - Ver la página en la caché de Google
    - Buscar páginas con contenidos similares
    - Navegación secuencial entre los resultados o yendo a una página de resultados

concreta (hasta la página 90 más o menos)

- Restringir la siguiente búsqueda a los resultados obtenidos



Actividad 5

Vaya a *Bing* (www.bing.com), haga una pregunta por cualquier término (p.e. “derecho al olvido”, “atentado en Francia”) y observe la página de resultados. Compárela con la página de resultados de Google. ¿Ofrece las mismas opciones? ¿Hay alguna opción en Ask que no esté en Google? ¿Cuál de los dos modelos de página de resultados le parece mejor?

Conclusiones

Conocer los aspectos estructurales y funcionales que hemos examinado aquí nos ayudará sin ninguna duda en nuestra relación con los motores de búsqueda. Conocer un poco mejor los fundamentos de los motores nos ayudará a llevar a cabo mejores búsquedas porque entendemos un poco mejor cómo "piensan" los buscadores y porqué determinados resultados están situados más arriba en el ranking que otros. Adicionalmente, si en nuestras responsabilidades entra publicar información en la web entenderemos mejor qué debemos hacer para nuestra página aparezca en los buscadores.

2. Búsqueda avanzada

Introducción

Afortunadamente, hasta un niño sabe utilizar la búsqueda simple. Los buscadores se han esforzado, siguiendo la estela de Google en ofrecer una opción de búsqueda simple realmente simple. Escribimos una palabra (o una frase) en la caja de búsqueda y hacemos Intro. Ya está. Es realmente importante que la búsqueda simple sea efectivamente muy fácil de llevar a cabo. Sin embargo, a veces, para búsquedas muy especializadas o para búsquedas en un contexto profesional o académico (en vez de ocio, por ejemplo) es necesario acudir a la búsqueda avanzada para obtener resultados de calidad.

Durante años, décadas en realidad, la búsqueda avanzada para recuperar información estuvo basado en la llamada lógica booleana, incluyendo a los motores de búsqueda de primera generación (AltaVista, por ejemplo). Como veremos, el conocimiento de la lógica booleana sigue siendo útil, puesto que la mayoría de los buscadores y bases de datos la utilizan de alguna forma; pero su importancia ha decaído en los últimos años debido a los modernos sistemas de cálculo de relevancia. No obstante, muchas veces lo que diferencia a un profesional de un aficionado en el mundo digital es precisamente este conocimiento. En esta unidad nos centraremos principalmente en las diversas variedades de la búsqueda booleana.

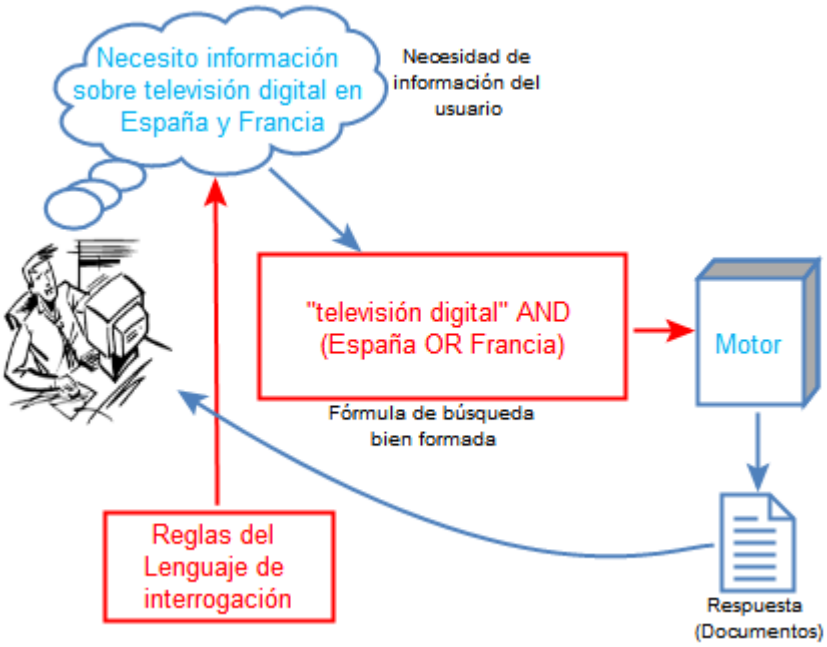
Lenguajes de interrogación

“Lenguaje de interrogación” (*query language*) es un término técnico heredado de las primeras bases de datos. Actualmente es trivial entrar uno o más términos en un buscador, lanzar una consulta y obtener con facilidad algunos resultados relevantes, de modo que referirse a esa operación como "lenguaje de interrogación" suena innecesariamente técnico (o peor, pedante).

Sin embargo, para que esa operación se haya vuelto tan simple ha debido darse un gran desarrollo conceptual y tecnológico. En el pasado, era necesario que el usuario explicitara de forma muy detallada qué quería encontrar usando lógica booleana (de la que nos ocuparemos enseguida), en qué campos quería encontrarlo, cómo quería que fuera la salida (es decir la respuesta), porqué campos quería que se ordenase la salida, etc. Un sistema de búsqueda era tan bueno como su lenguaje de interrogación. Aunque los documentos estuvieran bien indizados, si el usuario no podía expresar con máxima precisión qué es lo que quería, el sistema proporcionaba una gran cantidad de ruido.

Todo esto ha quedado minimizado actualmente debido al uso de sistemas de relevancia; sistemas que eran prácticamente inexistentes antes del 2000. Por todo ello, no es que los lenguajes de interrogación hayan dejado de existir o de tener importancia, es que su papel nos pasa simplemente desapercibido. Por tanto, este papel no ha variado, simplemente, los modernos sistemas de búsquedas hacen que pase desapercibido; pero como indicamos en la introducción, precisamente, en este conocimiento radica la diferencia entre el profesional y el aficionado.

Como es fácil deducir por lo que hemos expuesto el papel del lenguaje de interrogación es nada menos que transformar una necesidad de información que, por definición, es algo intangible, en una fórmula bien formada que un robot (un motor de búsqueda o una base de datos) pueda interpretar. La figura siguiente intenta ilustrar esta idea.



Como podemos ver, la simple idea es la siguiente: el primer paso siempre consiste en que un usuario tiene una necesidad de información; es decir aquellos que algunos teóricos anglosajones (Belkin, 1982)denominan un "*ASK*" (por *Anomalous State of Knowledge* [1]).

A partir de aquí, se considera que el lenguaje natural de los usuarios no siempre es utilizable directamente como lenguaje de interrogación, que debe cumplir algunas reglas. Por tanto, el segundo paso consiste en que el usuario directamente si tiene un buen conocimiento del lenguaje de interrogación o por medio de un intermediario (un documentalista o un asistente del propio sistema de búsqueda) transforma esa necesidad en una fórmula bien formada (FBF) del lenguaje de interrogación, en este caso, en una fórmula booleana que combina tres términos de búsqueda y dos operadores booleanos (AND, OR).

Esta FBF propia de su lenguaje de interrogación propio es la que puede entender el motor de búsqueda, en este caso suponemos que el usuario quiere documentos que tengan la frase "televisión digital" siempre que además (AND) tengan o bien la palabra España, o bien la palabra Francia o ambas (OR). La respuesta se supone que serán documentos que cumplan la condición

[1]Se supone que un estado mental en el que carecemos de información, lo percibimos como anómalo y queremos solucionarlo.

expresada de forma precisa por el usuario usando una sintaxis concreta del lenguaje de búsqueda del sistema.

Tabla 1: Tipos de búsqueda más usuales

Tipo	*Ejemplos*	*Explicación*
Simple	cine	Busca documentos que contengan la cadena de caracteres.
Frase	"john ford"	Fuerza al sistema para que la cadena de caracteres buscada sea exactamente como aparece entre las comillas. En este caso, primero la palabra john, después un espacio en blanco y después la palabra ford sin nada más entre ellas. Evita las falsas coordinaciones. Por ejemplo, la página de una empresa llamada John Smith que publica información de los coches de la marca Ford.
Booleana simple	televisión AND calidad O bien: televisión OR tv televisión NOT tv3	Principales operadores: AND, OR, NOT. **AND**: exige que ambas palabras estén presentes en el documento. En algunos motores de búsqueda, como Google, un espacio en blanco equivale siempre a un AND. **OR**: al menos una de las palabras debe estar presente; pueden estar presentes las dos. **NOT**: la primera palabra debe estar presente, siempre que no aparezca la segunda palabra. En este caso, todos los documentos que tengan la palabra cine, siempre que no tengan la palabra novela.
Booleana compleja	(televisión OR audiovisual) AND (calidad OR prestigio)	Combina tres o más palabras clave con dos o más operadores booleanos. Los paréntesis se pueden utilizar para delimitar de forma exacta el orden y el alcance de cada operador. En este caso: primero se crea al conjunto 1 con todos los documentos que tienen la palabra cine, la palabra tv o ambas. Después se crea el conjunto 2 con todos los documentos que tienen la palabra novela, la palabra teatro o ambas. Por último, se combinan los conjuntos 1 y 2 con el operador AND y se produce como respuesta el conjunto 3 que contiene el resultado final.

Además de las búsquedas anteriores, algunos motores (Google y Yahoo especialmente) ofrecen una gran variedad de posibilidades que a veces tiene poco que ver en realidad con la recuperación de información. Por ejemplo, el lenguaje de interrogación de Google actúa también como una calculadora aritmética si entramos expresiones de suma, resta, etc. (p.e. 9*670/8). Otras características se asemejan a lo que podríamos llamar un sistema de respuestas directas, es decir, no un sistema que busca documentos que contengan la respuesta, sino un

sistema que responde a la pregunta.

Por ejemplo, imagine que entramos la siguiente expresión en Google:

moneda de Jordania

y como respuesta tenemos:

Dinar jordano

en lugar de una lista de (solamente) documentos que tienen la expresión “moneda de Jordania”.

Actividad 6
Entre en esta página de Google: http://www.google.com/intl/es/help/features.html

Revise la lista de opciones y repase mentalmente cuántas de esas clases de búsquedas conocía usted (o no) previamente. Pruebe al menos una o dos de las que le produzcan mayor curiosidad o conozca menos.

Tabla 2: Búsqueda parametrizada. Variaciones sobre las búsquedas más usuales

Variación	*Ejemplo*	*Explicación*
Genérica	cine	Busca la cadena de caracteres en cualquier lugar del documento, en documentos de cualquier fecha, en todos los formatos de documento, etc.
Acotada	allintitle:cine filetype:pdf	Busca la cadena de caracteres en el campo título (allintitle:cine) y solamente en documentos con formato pdf (fileype:pdf). La sintaxis del ejemplo es la de Google. Otros buscadores pueden utilizar otras expresiones. Acotaciones principales: Por campos (p.e. título) Por fechas (p.e. Últimos 6 meses) Por tipo de documento (p.e. pdf) Por idioma (p.e. español) Por dominio (p.e: .com)

Tabla 3: Interrogación no booleana vs. booleana

Tipo de interrogación		*Significado lógico*
No booleana *(sin operadores explícitos)*	cine literatura niños adolescentes	Documentos que contengan todas y cada una de las palabras. No capta la necesidad real del usuario. Se pierden de feacto, por ejemplo, documentos que traten *solamente* de cine y de adolescentes. También los que traten *solamente* de cine y de niños, etc; puesto que pedimos exactamente que todas, es decir, las cuatro palabras estén presentes.
Booleana *(con operadores explícitos)*	(cine **OR** literatura) **AND** (niños **OR** adolescentes)	Capta exactamente la necesidad de información del usuario. Por ejemplo, un documento sobre niños y literatura será recuperado, aunque no hable ni de adolescentes ni de cine, etc., y esto es exactamente lo que necesita el usuario.

Como se puede ver, solamente la forma booleana capta de forma adecuada la semántica de esta necesidad de información, esto es, solamente la fórmula booleana es capaz de evitar tanto el ruido (información no deseada) como el silencio (información no recuperada).

El principal inconveniente es que el álgebra booleana es, de facto, anti intuitiva e intimidatoria para el usuario final. Por ejemplo, si revisamos la representación de la necesidad de información anterior:

cine **y** literatura **para** niños **y** adolescentes

la mayor parte de los usuarios que se hubieran visto obligados a traducirla a una expresión booleana lo hubieran hecho así:

cine **AND** literatura **AND** niños **AND** adolescentes

Por el simple procedimiento de traducir todos los **y** gramaticales en **AND** lógicos (incluida la expresión **para**), y que es obviamente incorrecta por las razones señaladas antes, pero que vamos recordar de nuevo: **no** estamos interesados **solo** en documentos que traten **todos** estos aspectos **en el mismo documento**; sino, literalmente en documentos que traten o bien de niños y cine, o bien de niños y literatura, o bien de adolescentes y cine, o bien de adolescentes y literatura, y ya sea por separado o conjuntamente.

Conclusiones

En el contexto de las bases de datos así como en base de datos de imágenes y video, el uso de búsquedas booleanas es imprescindible, dado que las referencias y las imágenes tienen poco texto comparado con el documento completo. Además, las bases de datos referenciales no suelen utilizar ningún tipo de ordenación por relevancia.

En las bases de datos de texto completo, el uso de búsquedas booleanas puede ser útil en función de la complejidad de la consulta y la calidad relativa del sistema de relevancia de la base de datos.

En cambio, en motores de búsqueda, su utilidad ha quedado en entredicho debido a dos factores: en primer lugar gracias a los sistemas de cálculo de relevancia de los motores que han minimizado la necesidad de expresar de forma tan estricta las necesidades de información. En segundo lugar porque la enorme abundancia de información de cualquier tipo hace cada vez menos importante la exhaustividad y más importante la precisión.

Es decir, al usuario típico de un motor que busca información por motivos de ocio o por simple curiosidad no le importa perder documentos relevantes mientras que los 10 o 20 primeros, que son los que estará dispuesto a ver como respuesta a su búsqueda, sean relevantes y solucionen (mejor o peor) su problema de información. La alternativa, en el contexto de una sobre saturación informativa, es decir, dedicar tiempo a aprender las bases de la lógica booleana y dedicar tiempo a planear cada búsqueda no es muy atractivo. Este es el motivo por el cual parece haber perdido peso no solamente el uso de las búsquedas booleanas explícitas (recordemos que toda búsqueda de dos o más términos lleva un AND booleano implícito), sino la misma necesidad de aprender la sintaxis de un lenguaje de interrogación o incluso la necesidad de acudir al sistema avanzado de búsqueda.

Los profesionales, en cambio, deben conocer bien estos aspectos, porque en alguna ocasión podrán resolver un problema de negocio o un problema crítico que ponga en cuestión su prestigio profesional gracias a un buen conocimiento de las características completas de los lenguajes de interrogación.

Además, como ya hemos señalado saber utilizar de forma adecuada un lenguaje de interrogación, en particular, el conocimiento de las búsquedas booleanas, es imprescindible en contextos donde no existe el documento completo (base de datos referenciales), donde el documento no es textual (bases de datos multimedia) o donde no pueden aplicarse ordenaciones de relevancia basadas en enlaces (bases de datos de texto completo).

3. Evaluación de sitios web

Finalmente, tanto los periodistas como el profesional de la comunicación están obligados a verificar sus fuentes, y esto es así tanto si por "fuentes" entendemos personas como sitios web (o cualquier clase de recursos digitales), como es el caso de esta Unidad.

3.1. Dimensiones de la calidad

Desde el punto de vista de la documentación periodística, la calidad de un sitio web presenta al menos dos parámetros básicos

1. Contenido
2. Credibilidad

3.1.1. Contenido

Los indicadores que nos ayudan a determinar la calidad de los contenidos de un sitio son, entre otros los siguientes:

1. Rigor. Se refiere al cuidado con el que ha sido preparada la información. No siempre podemos estar seguros respecto a esto, y con frecuencia debemos guiarnos por indicios más o menos (in)directos. Algunos de estos indicios son: la mención de otras fuentes, la mención de datos concretos (fechas, autores, estadísticas), la ausencia de errores factuales, la ausencia de juicios de valor gratuitos, la separación cuidadosa entre opinión e información, etc., por no mencionar la ausencia de errores ortográficos, etc.

Un buen ejemplo de esto suelen ser sitios web de gran relieve en el mundo de las políticas económicas, como el sitio web del Banco Mundial (www.bancomundial.org) o el sitio de la OCDE (http://www.oecd.org).

2. *Exhaustividad*. Este indicador se refiere al grado en el cual las informaciones se presentan de modo más o menos completo. Por ejemplo, un dato bibliográfico puede consistir solamente en un título y un nombre de autor, o incorporar también la fecha de edición, el lugar de edición, el nombre de la editorial, el número de páginas de la obra y el número de ISBN. Un directorio de empresas puede dar un nombre y una ciudad para cada empresa o la ficha completa de la misma, incluyendo el nombre de los miembros del consejo de administración; una biografía puede dedicar un párrafo a cada personaje o el equivalente a varias páginas de información, etc.

Un excelente ejemplo de exhaustividad en la mayor parte de sus informaciones la encontramos en la base de datos de cine IMDB (www.imdb.com).

3. *Actualización*. Señala la frecuencia o la fecha de la última actualización. Es evidente que, en general y a igualdad de otros factores, cuanto más actualizada sea una información, mayor valor tiene. Si consultamos un atlas geográfico de centro Europa desearemos, con toda seguridad, que sea lo más actualizado posible.

En general, los sitios web de medios diarios, son ahora buenos ejemplos de una política adecuada de actualización, con varias ediciones diarias.

4. *Edición*. Este indicador se refiere al grado, tanto explícito como implícito, y más o menos evidente según los casos, en el que la información ha sido tratada, revisada, completada, editada en una palabra, antes de ser publicada. Ofrece evidencias de edición una web que presenta un título en cada página, un nombre de autor (o una indicación de autoría) y una fecha de creación, por no mencionar, una vez más, la ausencia de errores ortográficos o la simple presentación descuidada. Las menciones de autor, la titulación y la datación deben repetirse en todas las secciones de la web si procede, es decir, si la naturaleza del recurso lo exige. Por ejemplo, la web de una publicación periódica deberá indicar no solamente los datos del organismo editor, el nombre del director o directora de la publicación, etc., sino que deberá presentar esos datos para cada artículo de la publicación, etc.

Desde hace algún tiempo, los proyectos de la organización Wikimedia, entre los que se incluye la Wikipedia son un buen ejemplo de edición cuidada, donde varios editores supervisan que los artículos de la wikipedia dispongan de datos verificados. Incluso cuando no es así, se señala la carencia de ellos para alertar al lector.

5. Sistematización. El tratamiento sistemático proporciona una gran calidad a la información, por eso debe considerarse un indicador de primera magnitud. En el punto anterior hemos mencionado un factor de sistematización, pero cada tipo de recurso tiene los suyos propios. Por ejemplo, en un directorio hay sistematización si cada web está descrita de la misma forma. Hay sistematización en una base de datos cuyos registros poseen todos la misma estructura básica, etc.

Las bases de datos que hemos revisado en esta Unidad cumplen todas con esta característica. Pero los principales diarios en línea también, ya que procuran presentar las noticias siempre con un tratamiento uniforme.

6. Interés intrínseco. Aunque sea una obviedad, hay que recordar que existen temas completamente triviales, y muchas webs dedicadas a ellos. De hecho, la trivialidad y el esnobismo parecen ejercer una atracción fatal en el mundo de la WWW. Dicho de otro modo, debe considerarse el interés intrínseco de la información por encima de consideraciones de estética o de esnobismo. En este sentido, otro motivo de sorpresa

es la enorme atención que algunas webs de contenidos voluntariamente triviales merecen en los medios de comunicación.

Una vez más, la Wikipedia o cualquier sitio de los que hemos comentado en esta Unidad cumplen este criterio.

7. Originalidad. Este parámetro se refiere al grado en el cual un recurso ofrece informaciones únicas o exclusivas en algún sentido. Muchas webs ofrecen informaciones relacionadas con temas de actualidad, pero solamente algunas webs ofrecerán, además, tratamientos en profundidad de tales temas o recursos complementarios, como gráficos, mapas, infografías, datos estadísticos, puntos de vista opuestos, etc. Muchas webs ofrecen listas de recursos digitales, pero solamente algunas evalúan esos recursos, etc. Este indicador mide, en definitiva, el grado en el cual un recurso ofrece algo que únicamente puede encontrarse en el mismo.

3.1.2. Credibilidad

Los indicadores anteriores son directamente observables o medibles y, en todo caso, relativamente fáciles de determinar.

El cambio, la credibilidad se refiere a una valoración donde, además de un inevitable componente subjetivo por parte del analista, tenemos que contar con que los responsables del sitio suelen estar interesados en parecer creíbles, independientemente de que lo sean o no.

Determinar la credibilidad de un sitio puede requerir indagaciones indirectas, así como efectuar inferencias sin garantía de acierto al cien por cien. Por último, la indagación puede no aportar ningún dato concluyente. Dicho de otro modo, mientras que siempre podremos saber o decir algo sobre el contenido, puede que no siempre podemos llegar a conclusiones seguras sobre la credibilidad.

No obstante, es éste último (la credibilidad) el aspecto que más debe preocupar en el contexto del trabajo periodístico por razones evidentes e independientes de su facilidad o dificultad de determinación.

En todo caso algunos indicadores relativamente fáciles de determinar son los siguientes:

Por su parte, el parámetro de la autoría puede evaluarse a través de tres aspectos o indicadores diferentes (aunque muy relacionados entre sí, claro está):

1. ***Página de créditos***. La existencia misma de declaraciones explícitas de autoría es ya una primera indicación de credibilidad. Sorprende el elevado número de webs que carecen de indicaciones claras sobre a quién cabe atribuir la paternidad, la propiedad o la responsabilidad intelectual de la misma. Por eso, de entrada, es ya un indicador de calidad la simple presencia de esa información bajo apartados denominados *Créditos, Quiénes somos*, *About Us*, etc. Una web bien concebida debe tener un apartado donde se explique sin ambigüedad y con el máximo detalle posible: a quién o a quienes debe atribuirse la responsabilidad del recurso; quién o quiénes son los editores o distribuidores del recurso. Esta declaración debería ir acompañada de datos complementarios sobre los autores y los editores: institución donde trabajan; dirección postal de los editores, etc. Por el contrario, una web sin tales indicaciones debe considerarse, de entrada, poco solvente.

   Un sitio que puede ser un buen ejemplo de cómo resuelve este apartado de forma ejemplar es la página de créditos de la televisión pública de Estados Unidos: PBS (www.pbs.org).

3. ***Solvencia o adecuación de la fuente***. La solvencia del autor y/o de la institución que ha creado, producido o editado el recurso. Es evidente que no merece el mismo crédito una web sobre salud producida por un doctor o una doctora en Medicina

que por un curandero, para poner un ejemplo de los llamados "de manual". Por tanto, un indicador de calidad en este sentido será la adecuación curricular del autor al tema tratado. Por parecido razonamiento, es evidente que no podemos otorgar el mismo crédito a un informe sobre política cultural publicado en la web de la Unesco que uno publicado en la web de un grupo que haga llamamientos al exterminio de media humanidad. Por tanto, respecto a la figura del editor o distribuidor, su solvencia pública será otro indicador de importancia capital. Como regla general, es evidente que existen organismos que ofrecen mayores garantías, de entrada, que otros. Por ejemplo, siempre suelen ser un buen indicador de calidad el que un recurso haya sido publicado por una universidad, o por una editorial de prestigio, etc.

Por ejemplo, la web del Banco Mundial o de la OCDE, ya indicadas anteriormente son buenos ejemplos de fuentes solventes por lo que hace a datos e informaciones económicas.

Ahora bien, ¿qué sucede cuando no existe en el sitio una indicación clara de autoría? O, peor aún, ¿qué sucede cuando tenemos sospechas sobre la autoría o la identidad declarada del sitio?

Ante la situación anterior, tenemos una pequeña batería de recursos que nos proporciona la propia web para intentar determinar "quién está detrás" de un sitio determinado.

Antes de presentar estos recursos, no obstante, una observación. Lo más operativo cuando nos encontramos ante un sitio del que no estamos seguros de su fiabilidad o credibilidad y su página de créditos no nos ayuda a despejar dudas consiste simplemente en ignorarlo y buscar otra fuente. En la web, por suerte, lo que *abunda* especialmente es la información, así que en general siempre tendremos alternativas donde escoger si una fuente no merece nuestra confianza.

Por tanto, lo que sigue solamente tiene sentido si, por alguna razón, y a pesar de todo estamos especialmente interesados en saber más sobre un sitio. Las herramientas que podemos utilizar son las siguientes (ninguna de ellas, sin embargo es concluyente):

1. ***Alexa.*** Este directorio (www.alexa.com) proporciona información muy importante sobre cualquier sitio. Nos indica: (i) el tráfico del mismo (el número de visitas que recibe); (ii) otros sitios que enlazan al sitio analizado; (iii) comentarios o revisiones sobre el sitio (si existen) y otras indicaciones estadísticas muy importantes sobre el sitio, permitiendo incluso hacer comparaciones entre varios sitios. Es muy importante el perfil que Alexa nos proporciona así de un sitio. Si apenas recibe visitas (o al contrario), o si quienes lo enlazan tienen un sesgo determinado, etc., en conjunto podremos hacernos una idea bastante determinada de la credibilidad del sitio. Naturalmente, no es un sistema infalible. En general, sirve más para confirmar la credibilidad de un

sitio que para lo contrario, porque sitios poco visitados y poco enlazados no aparecen en su directorio. No obstante, la ausencia de un sitio en Alexa también puede ser (no necesariamente) un elemento de advertencia. Siempre puede ayudar saber qué sitios enlazan al sitio considerado. Por ejemplo, un sitio sobre salud que recibe enlaces principalmente de sitios vinculados con la parasicología o procedentes de una secta, etc., naturalmente no es un buen indicio (tampoco es infalible, un sitio puede ser enlazado sin permiso de los responsables del mismo).

2. ***Servicios Whois***. Existe una serie de sitios que permiten lanzar consultas a los registros oficiales de los propietarios de los dominios. Un ejemplo es Whois.net (http://www.whois.net/). No son datos concluyentes, pero típicamente podemos saber el nombre de la empresa o de la persona que compró el dominio y sus datos básicos de contacto. No obstante, a veces están ocultos.

3. ***Open Site Explorer*** (https://moz.com/researchtools/ose/). Es un buscador especializado que analiza la URL que entremos y nos indica informaciones básicas sobre el sitio, algunas son en realidad muy similares a las que proporciona Alexa (qué otros sitios enlazan al sitio determinado, p.e.) pero siempre nos servirá para ampliar o comparar resultados y en teoría nos proporciona información de cualquier sitio aunque no forme parte del top 100 mil como en el caso de Alexa. **Alternativa: Ahref** (https://ahrefs.com/).

4. ***Búsquedas en la web***. Naturalmente, siempre tenemos la opción de entrar el nombre y/o la URL del sitio en la caja de búsqueda de Google y ver qué dice el resto de la web (si es que dice algo) sobre el sitio en cuestión. Si no dice nada, ya es de por sí motivo de sospecha.

5. ***Uso de navegadores de última generación***. Aunque parezca una trivialidad un periodista siempre debería utilizar navegadores de última generación, o al menos de versiones muy recientes. La razón es simple: cada nueva generación de los principales navegadores (IE, Firefox, Opera) añade mejores capacidades para detectar sitios fraudulentos o que usurpan la identidad de sitios reales, etc.

3.4. Conclusiones

Determinar la calidad de un sitio web puede ser absolutamente trivial. No hace mucho análisis para observa la enorme calidad de sitios como los del *New York Times*, *The Guardian*, *PBS*, *El*

País, etc. Otras veces, tropezamos con páginas infames que inmediatamente nos advierten sobre su falta total de credibilidad.

Sin embargo, el problema nunca son este tipo de sitios que se sitúan en los extremos, sino aquellos que encontramos a través de una búsqueda en la web, de los que antes no sabíamos nada y presentan una apariencia como mínimo correcta. En estos sitios podemos encontrar información aparentemente útil y sin embargo tener dudas fundadas sobre su solvencia.

Si observamos unas pautas elementales, en realidad de sentido común, podemos evitar ser embaucados por sitios que no merecen nuestra confianza. En este apartado hemos proporcionado además algunas pautas que podemos seguir si, por algún motivo, es importante para nosotros saber más sobre un sitio y asegurarnos de su credibilidad o solvencia sobre los temas que trata.

Actividad 7

Haga una búsqueda en Google (o en cualquier otro buscador) sobre un tema de su interés. Seleccione los tres primeros sitios de la primera página de resultados de Google y compare las respectivas páginas de créditos de cada uno de los sitios. Haga capturas de pantalla comentadas.

6. Referencias

Abadal, Ernest; Codina, Lluís. 2005. *Bases de datos documentales: características, funciones y método*. Madrid: Síntesis, 220 p.

Cabezas, Álvaro; Daniel Torres; Emilio Delgado. 2009. "Ciencia 2.0: Herramientas e implicaciones para la actividad investigadora". *El Profesional De La Información* 18, (1): 72-9.

Codina, Lluís; Palma, M. del Valle; CID, Pilar. 2006. "Búsqueda Multimedia en la Web". Barcelona: COBDC. *IV Jornadas de Documentación Audiovisual*. Barcelona, Mayo 2006. Acceso: http://www.codc.org/jornades/10JCD/docaudio.html. Acceso alternativo: http://www.lluiscodina.com/multimedia2006.ppt)

Codina, Lluís. *Web 2.0* (diagrama y directorio). 2009 Acceso: http://www.mindomo.com/view.htm?m=2671b7eeda744faab517aa258e9ad4e8

Codina, Lluís. *Ciencia 2.0*. (diagrama y directorio). 2009 Acceso: http://www.mindomo.com/view.htm?m=d4d1f77be0d04af0804c719038144de8

Foust, James C. 2009. *Online journalism: Principles and practice for the web*. 2nd edition ed. Scottsdale: Holcom Hataway.

Jiménez, Marina; Ortiz, Virginia. 2007. *Evaluación y calidad de sedes web*. Gijón: Trea, 270 pp.

Kovach, Bill, and Tom Rosnestiel. 2007. *The elements of journalism: What newspeople should know and the public should expect*. New York: Three Rivers.

López Yepes, Alfonso. 2007. “e-DocuInfo: un portal de documentación informativa para la formación, investigación y producción multimedia en línea”. *Boletín de la ANABAD*, pags. 49-70

López Yepes, Alfonso; Sánchez Jiménez, Rodrigo. 2007. “e-DocuInfo/TVDoc: portal de documentación informativa-canal temático iptv”. *El profesional de la información*, Vol. 16, Nº 5, pags. 456-467

López Yepes, José (coordinador). 2002. *Manual de Ciencias de la Documentación*. Madrid: Pirámide, 742 p.

Maciá, Fernando, and Javier Gosende. 2009. *Posicionamiento en buscadores*. Madrid: Anaya.

Moreiro, José Antonio (coord.) 2000. *Manual de documentación informativa*. Madrid: Cátedra, 458 p.

Forma de citación de este documento: [Lluís Codina, Rafael Pedraza. *Sistemas de Información Documental para profesionales de la Comunicación*. Barcelona: Universitat Pompeu Fabra, 2015 (25 p.)] Sitio web de uno de los autores: www.lluiscodina.com