**Algoritmo de Crawling**

**Introducción.-**

El rastreo web es el proceso mediante el cual reunimos páginas de la web para indexarlas y dar soporte a un motor de búsqueda. El objetivo del rastreo es reunir de forma rápida y eficiente el mayor número posible de páginas web útiles, junto con la estructura de enlaces que las interconecta.

Un Web Crawler es un sistema para la descarga masiva de páginas web. Los Web Crawlers son usados para varios propósitos. El más importante, es uno de los principales componentes de un motor de búsqueda, que son sistemas que ensamblan los cuerpos de las páginas web, los indexa y permite a los usuarios emitir consultas y encontrar las páginas web que coincidan con la consulta.

**Definiciones.-**

**¿Qué es el Crawling?**

El crawling o rastreo de sitios web es el recorrido que hace un pequeño bot de software (un crawler) para leer y analizar el código y contenido de una web, saltando de página en página a través de los enlaces que va encontrando.

En el caso del GoogleBot (el crawler de Google), este se encarga de rastrear y examinar nuestras webs, para posteriormente incorporarlas a su índice.

**Crawler.-**

Básicamente es un algoritmo que aprovecha la estructura de la red para moverse de página a página.

El algoritmo básico ejecutado por cualquier web crawler toma una lista de URLs "Semilla" como su entrada y ejecuta repetidamente los siguientes pasos:

* Se quita una URL de la lista de URLs, se determina su dirección IP, se descarga el documento correspondiente y se extraen los enlaces que contenga.
* Para cada uno de los enlaces extraídos, nos aseguramos de que es una URL absoluta y la añadimos a la lista de URLs que descargar, siempre que no haya sido encontrada antes.
* Las URLs de la lista de URLs se visitan recurrentemente según un sistema de prioridades. Esta prioridad puede implementarse mediante métricas de importancia, como aquellas basadas en la cuenta de enlaces que apuntan a la página (IB) o aquella basada en el peso relativo o "importancia" de cada página(IR, utilizada por Google).

Este es el algoritmo básico, que requiere un número de componentes funcionales:

* Un componente llamado frontera de Urls para almacenar las lista de Urls a descargar
* Un componente para resolver los nombres de host a direcciones IP
* Un componente para descargar documentos usando el protocolo HTTP
* Un componente para extraer enlaces de los documentos HTML
* Un componente para determinar si la URL ha sido encontrada antes

**Cómo funcionan los crawlers?**

En principio un crawler es como un bibliotecario. Busca información en el sitio web, la categoriza y la indexa antes de que venga analizada.

Las operaciones de este programa deben establecerse antes de que los crawler empiecen a analizar un sitio web. El crawler procesa estas órdenes automáticamente y, por lo tanto, el tipo de información que obtiene el crawler depende exclusivamente de las normas que se le hayan establecido.

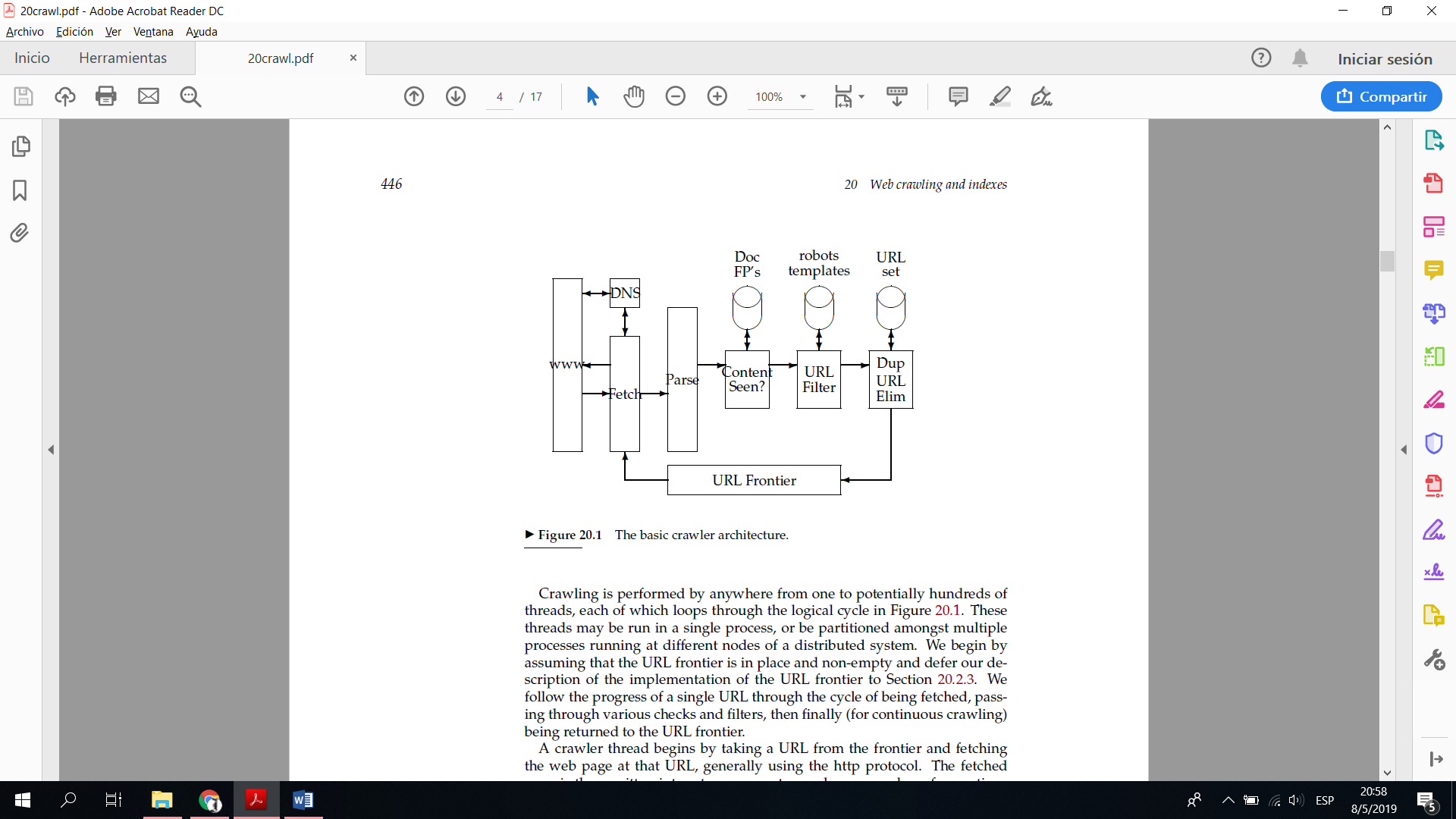
Existe un archivo llamado robots.txt donde le indicas cual sería el contenido que los buscadores deban tomar en cuenta. Con la información obtenida de tus sitios, los servidores crean un índice en su bases de datos.

### **Para qué sirve el Crawling**

El Crawling sirve para repasar todo el contenido y enlaces de una sitio web con el fin de indexar todo en los buscadores correspondientes. Es decir, sirve para que cada actualización o nuevo contenido que se publique en una página acabe apareciendo en Google y demás motores.

**Arquitectura básica del crawler.-**

1. URL frontier (frontera de URLs): contiene los URLs que aún no fueron recuperados por el crawler.
2. DMS resolution module (módulo de resolución DNS): determina el servidor web de donde se recuperará la página especificada por un URL.
3. Fetch module (módulo de recuperación): utiliza el protocolo http para recuperar la página web de una URL.
4. Parsing module: extrae el texto y el conjunto de links de una página web recuperada.
5. Tres módulos de eliminación: Uno determina si la pagina web recuperada tiene el mismo contenido de una URL ya revisada, el segundo modulo es usado para determinar si la URL obtenida debera ser excluida de la frontera; esto se determina en base a algunos tests, y el ultimo modulo determinan si un URL extraído de una página web ya se encuentra en la frontera de URLs.



**Pseudocódigo de un algoritmo de crawling.-**

Encolar(Cola\_URL, URL\_Inicio);

Mientras (No( Vacia (Cola\_URL)){

url=Desencolar(Cola\_URL);

pagina=minarpagina(url);

Encolar(paginas\_recogidas,(url,pagina));

lista\_url=extraer\_urls(pagina);

para cada u en lista\_url

Encolar(enlaces,(url,u));

Si [u no en cola\_url ]Y[(u,-) no en paginas\_recogidas]

Encolar(Cola\_URL,u);

Reordenar\_Cola(Cola\_URL);

**Algoritmos**

***Primero el mejor – Ingenuo***

Usa similaridad por coseno. Se calcula el valor de similaridad de la página de la cual extrae los URLs con la consulta:

https://lh6.googleusercontent.com/thm0DY9TMvkMkPx-7caOcwfhx6pksg3BP4LE-bGmx_tlSMgCP9OMwkk90vVuRgQkMT1WsTOPB-AxJfF5qOmttGplVPUhTzILMFdHDKcPTiFL2HU21hTR774lgCbBKAVWzukfiKEP

Donde q es la consulta, p es la página de la cual extraje los URLs, Vq y Vp son las representaciones vectoriales de las frecuencias de los términos, para la consulta y la página respectivamente. sim(q,p) es puntaje que se le da a los nuevos URLs.

***Crawlers enfocados -> FER***

La idea básica de estos crawlers es clasificar las páginas en una taxonomía de temas.

El usuario puede ir agregando nuevas categorías e ir marcando algunas como “buenas”.

Además se deben dar unos URLs ejemplo que el crawler usa para construir un clasificador capaz de encontrar la probabilidad de que una página pertenezca a cierta categoría.

Se calcula un puntaje de relevancia de acuerdo a que tan buena es la categoría en la cual cae la página.

***Crawlers enfocados en el contexto***

Es un poco más elaborado que el anterior, usa también clasificadores, pero no descarta inmediatamente las páginas que no le parecen importantes.

Espera unos links más, con la esperanza de que haya algo importante unos pasos más allá en la red.

**Ejemplo de aplicación.-**

## **¿Cómo realiza Google el crawling de las webs?**

En primer lugar, Google deberá conocer de la existencia de nuestra sitio web y de su disponibilidad para que el crawler realice el proceso de rastreo e indexación de la misma. Para ello, contamos con diferentes opciones, siendo las más habituales la creación de un enlace externo hacia el sitio web en cuestión o el alta en la plataforma Search Console, entre otras.

Pasado este punto, el crawler de Google empezará el proceso de rastreo del portal, accediendo a todas las páginas a través de los distintos enlaces internos que hayamos creado. También puede beber de otras fuentes para encontrar una página o sección de nuestra web, como su existencia en un archivo Sitemap que hayamos dado de alta en Search Console.

**¿Cómo funciona realmente?**

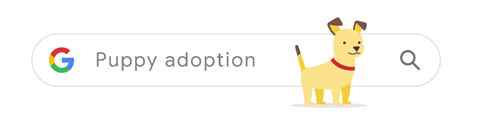
El motor de búsqueda de Google (Chrome) está en constante crecimiento. Actualmente recoge más de 30 millones de millones de páginas, maneja alrededor de 100 millones de gigabytes, recibe unos 100.000 millones de búsquedas mensuales.

Recalcular y ordenar este material se produce a través de un procedimiento llamado “[crawling](http://www.googleguide.com/google_works.html)”, que recorre las diferentes páginas y las indexa en un periodo de tiempo de 2 días. De esta forma, los propietarios de los portales pueden decidir si su contenido se rastrea o no. Google ordena las páginas en función del contenido y otros factores mediante una serie de fórmulas matemáticas que ofrecen *“el mejor resultado”.*

Esos algoritmos ponen manos a la obra al motor de búsqueda para traducir lo que los usuarios pretenden encontrar. Lo consiguen mediante diversas funcionalidades como la proporción de alternativas de búsqueda (*entre otros muchos*). Basado en esos parámetros se devuelve un ranking de resultados usando más de 200.000 variables.

**¿Cómo funciona la búsqueda de Google?**

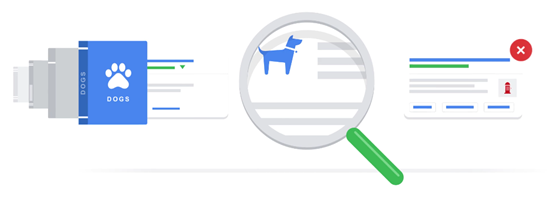
Cada vez que se realiza una búsqueda, hay miles, a veces millones, de páginas web con información útil. Google se da cuenta de qué resultados mostrar mucho antes que el usuario escriba.



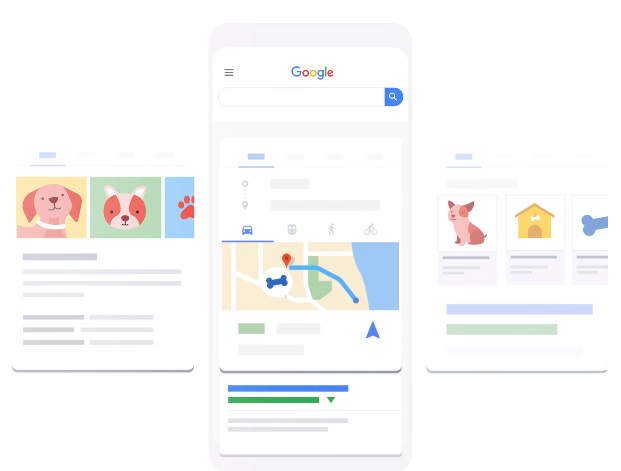
Incluso antes de realizar una búsqueda, Google organiza información sobre páginas web en su propio índice de búsqueda. El índice es como una biblioteca, excepto que contiene más información que en todas las bibliotecas del mundo juntas.



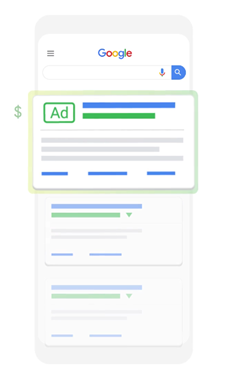
En una fracción de segundo, los algoritmos de búsqueda de Google clasifican cientos de miles de millones de páginas web en nuestro índice de búsqueda para encontrar los resultados más relevantes y útiles para lo que está buscando.



Para ayudar al usuario a encontrar lo que está buscando rápidamente, Google proporciona resultados en muchos formatos útiles. Ya sea que se presente como un mapa con direcciones, imágenes, videos o historias, Google está evolucionando constantemente con nuevas formas de presentar información.



Si bien los anunciantes pueden pagar para que se muestren en secciones claramente marcadas de la página, nadie puede comprar una mejor ubicación en los resultados de búsqueda.



**¿Y qué factores pueden afectar negativamente a nuestro Crawl Budget (tiempo que los buscadores dan a sus** [**crawlers**](https://neoattack.com/neowiki/crawlers/) **para rastrear cada página web)?**

Todos aquellos problemas o limitaciones que hagan gastar más tiempo y recursos al bot de Google, como por ejemplo:

* Un alto tiempo de respuesta del servidor
* Velocidad de carga de recursos excesivos
* Presencia de errores de código 4xx o 5xx
* Páginas aisladas o de difícil acceso

**Ventajas**

* Se puede obtener una gran cantidad de datos gracias al proceso de crawling.
* La codificación de un crawler puede dar muchos beneficios durante un largo tiempo.
* Si una página web es incluida en algún índice, lista o base de datos, puede obtener tráfico adicional.
* Ya que internet es la fuente de datos más grande y perpetua de internet, cualquier compañía desearía tener un puesto privilegiado con suficiente tráfico de visitantes.
* Se puede utilizar los datos obtenidos del crawler para crear nuevas aplicaciones que agreguen, indexen o proporcionen información.
* Proporciona mayor accesibilidad a la información.
* Ayuda a categorizar y organizar los grandes volúmenes de datos.

**Desventajas**

* Su tráfico puede ser identificado como abusivo o sospechoso y bloqueado.
* Puede ser limitado por el ancho de banda, de procesamiento o de almacenamiento.
* Se debe tener en cuenta los aspectos legales, como ser la privacidad de datos.
* En el intervalo entre actualizaciones los buscadores pueden devolver enlaces inválidos (2 - 9 %).
* No se indexan páginas generadas dinámicamente o protegidas con contraseñas.
* Los crawlers pueden tener problemas para indexar páginas con frames o mapas de imágenes.

**Conclusión.-**

El crawling es un componente muy importante en los motores de búsquedas, que nos ayuda a leer y analizar el contenido de las páginas, aprovechando toda la estructura de la red.