

Minería en la web: búsqueda, crawling y scraping

Sistemas de Recuperación de Información

Lic. Carlos León González
Dra.C. Lucina García Hernández

Facultad de Matemática y Computación
Universidad de La Habana

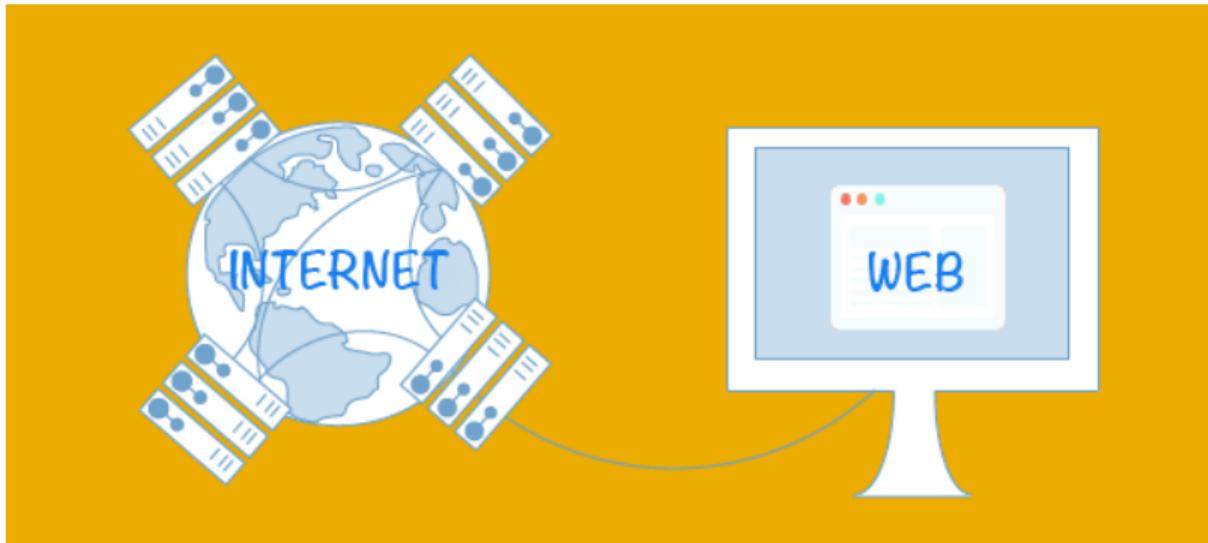
17 de junio de 2024

Objetivos

- Identificar y caracterizar las distintas etapas del desarrollo de la Web.
- Abordar los distintos componentes de un motor de Recuperación de Información en la Web.
- Definir las funciones del crawler y el scraper.

Duda

¿Representan lo mismo?



Tomado de <https://ladiferenciaentre.info/internet-web/>

Diferencias entre Internet y Web

Internet	Comparativa	Web
Infraestructura de comunicación	¿Qué es?	Conjunto de documentos conectados a través de hipertexto.
ARPANET fue la primera red de computadoras. Creada por el Departamento de Defensa de EE.UU. en 1968.	Creación	La primera propuesta de la web fue presentada por Tim Berners-Lee en el CERN en 1989.
Funciona con el protocolo TCP/IP que divide los datos en paquetes y los une en el destino	¿Cómo funciona?	Los documentos se codifican en HTML y se envía a través del protocolo HTTP

Diferencias entre Internet y Web

Internet	Comparativa	Web
<ul style="list-style-type: none">■ Descentralizada■ Abierta■ Escalable■ Diversidad de protocolos	Ventajas	<ul style="list-style-type: none">■ Acceso mediante hipervínculos■ Navegación por páginas web■ Interactividad
<ul style="list-style-type: none">■ Acceso no garantizado■ Vulnerabilidad a ataques■ Congestión■ Falta de control	Desventajas	<ul style="list-style-type: none">■ Dependencia de Internet■ Requiere software específico■ Contenido no siempre veraz■ Limitaciones de accesibilidad

Pero ...

¿Siempre la Web ha sido igual?
¿Ha sufrido algún cambio?

Pero ...

La Web ha sufrido cambios desde su surgimiento a finales de la década de 1980 y se ha dividido en distintas eras.

Era: Web 1.0

Características:

- Comunicación unidireccional
- Interacción limitada del usuario
- Sitios web estáticos
- Falta de interacción social
- Ausencia de contenido dinámico



Tomado de <https://laeradigitaltierno.home.blog/2019/01/17/3-1-evolucion-de-la-web/>

Web 1.5

Puede considerarse que la incorporación de las bases de datos en los sitios web estáticos de la **Web 1.0** es un paso superior, dándose la **Web 1.5**.

El uso de las bases de datos por los sitios web tenía como objetivo atraer más usuarios en la red.

Web 1.5

Puede considerarse que la incorporación de las bases de datos en los sitios web estáticos de la **Web 1.0** es un paso superior, dándose la **Web 1.5**.

El uso de las bases de datos por los sitios web tenía como objetivo atraer más usuarios en la red.

La **Web 1** se conoce también como **Web de solo lectura** o **Web estática**.

Era: Web 2.0

En 2004 Tim O'Reilly introduce el concepto de Web 2.0, definiéndolo como:

Un conjunto de aplicaciones donde el usuario tiene el control y se caracterizan por estar basadas en la inteligencia colectiva y el uso de servicios interactivos en red.

Era: Web 2.0

Características:

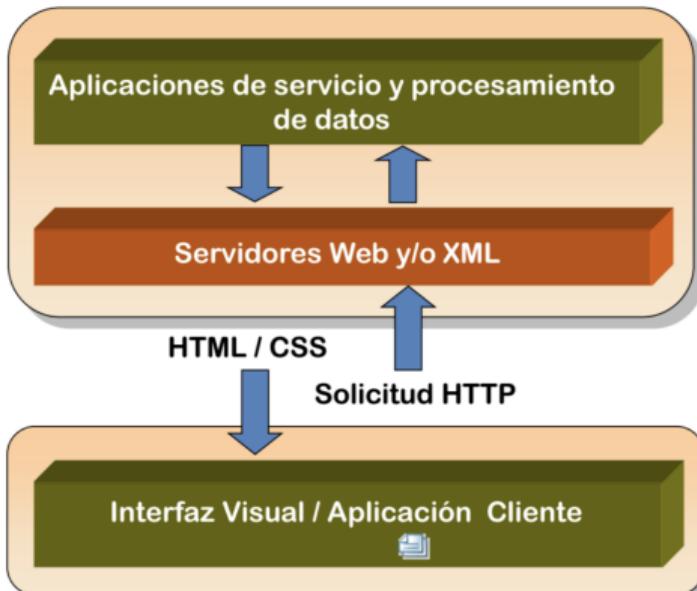
- Contenido generado por el usuario
- Interactividad y participación
- Uso de plataformas sociales
- Aplicaciones enriquecidas de Internet (AJAX, JavaScript, Flash)
- Personalización y adaptación



Avances en las aplicaciones web

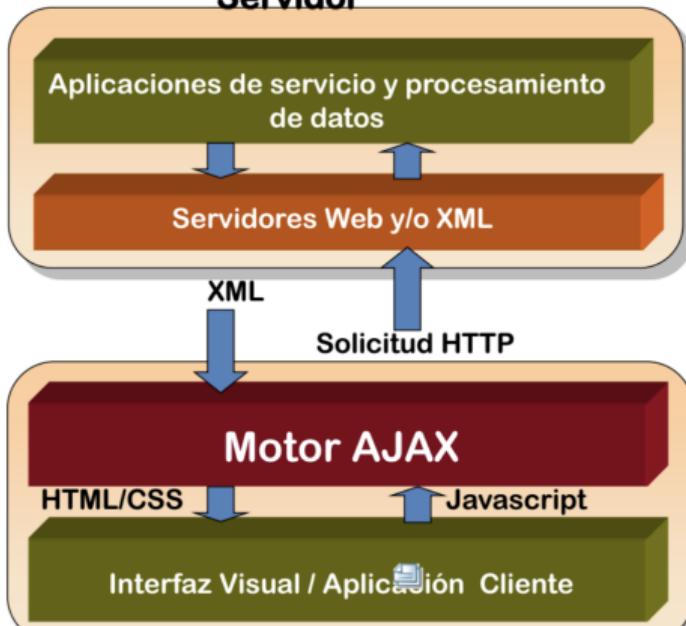
Web 1

Servidor



Web 2

Servidor



Si no apareces

en Google,

no existes.



Web 2.5

La incorporación de las redes sociales, la popularización de los usuarios y las entidades y la comunidad virtual generadas crean una transición a una era de la web con mayor especialización: **Web 2.5**.



Tomado de <https://www.nunify.com/blogs/build-virtual-communities-around-virtual-events/>

Comunidad virtual

Conjunto de individuos que con un mismo objetivo o propósito se unen para coincidir en una Red Social Virtual apoyándose en tecnologías que permiten realizar esta relación de forma virtual.

Web 2

La **Web 2** se conoce también como **Web de escritura-lectura** o **Web social**.

Problemas de la Web 2

- Diseñada para el uso humano
- Crecimiento por segundo del volumen de información disponible
- Evoluciona como un repositorio caótico
- Diseño no adecuado para la publicación y recuperación de información de manera “organizada”
- Resultados irrelevantes en las búsquedas
- Web totalmente sintáctica

Problemas de la Web 2

- Diseñada para el uso humano
- Crecimiento por segundo del volumen de información disponible
- Evoluciona como un repositorio caótico
- Diseño no adecuado para la publicación y recuperación de información de manera “organizada”
- Resultados irrelevantes en las búsquedas
- Web totalmente sintáctica

¿Qué hacer para intentar dar significado dentro de la Web?

Web semántica

En 2001 se propuso por el propio creador de la web, Tim Berners-Lee, un nuevo rasgo para añadir en la web, siendo:

Web semántica

Es una extensión de la web actual (Web 2.0) en la cual la información recibe un significado bien definido, permitiendo a los computadores y las personas trabajar en cooperación de mejor forma.

Su objetivo es que la web sea más comprensible para los agentes de búsqueda en términos de relevancia y precisión.

Web 3

- Descentralización y tecnología Blockchain
- Control del usuario y propiedad de los datos
- Privacidad y seguridad mejoradas
- Interoperabilidad e integración perfecta
- Integración inteligente de la Web y la IA
- Espacios tridimensionales



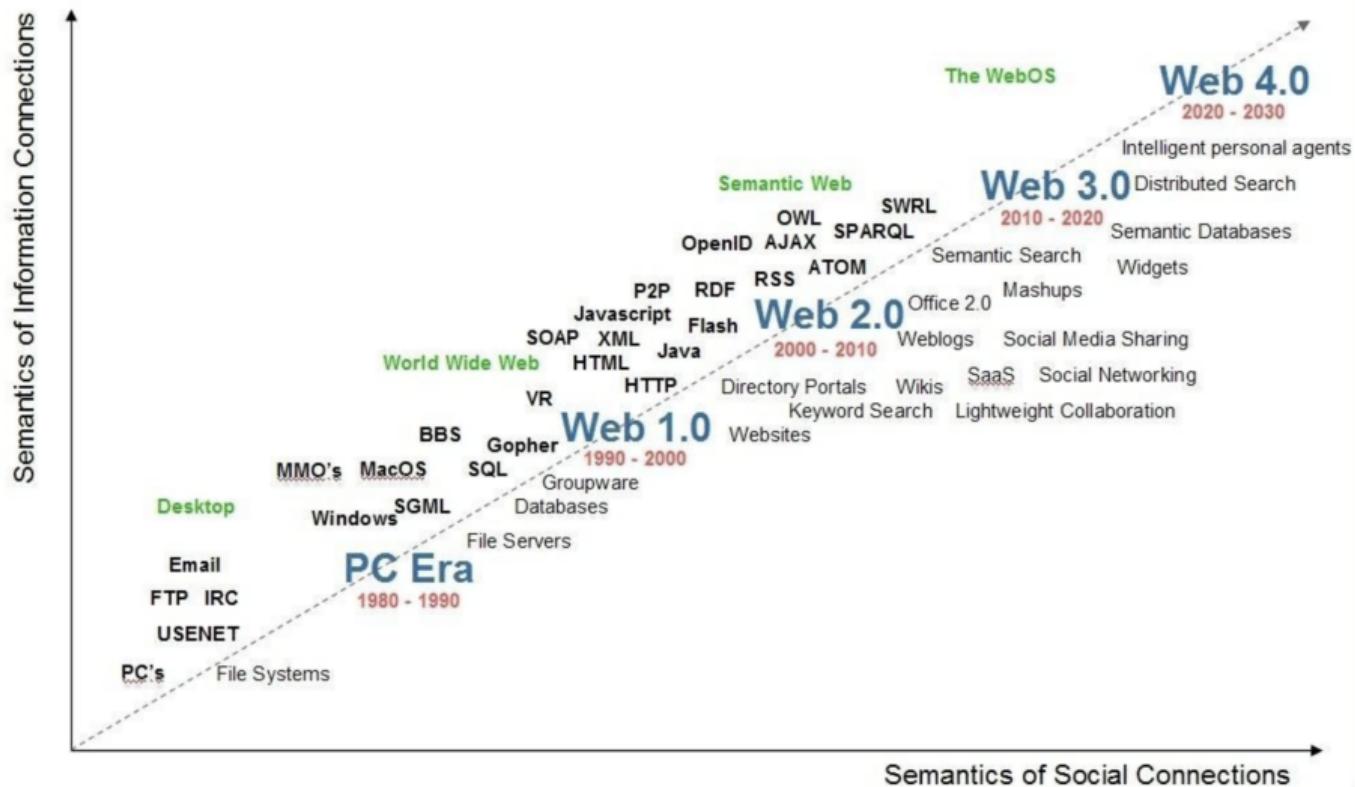
Tomado de <https://laeradigitaltierno.home.blog/2019/01/17/3-1-evolucion-de-la-web/>

Posterior a Web 3 ...



Tomado de <https://www.linkedin.com/pulse/web-4-0-explained-brief-agiledistrict/>

Evolución (proyección) de la web



Tomado de <https://www.linkedin.com/pulse/20141119114952-62080297-evolution-of-web-development/>

Extra para ampliar el conocimiento y posibles análisis

El sitio <https://web.archive.org> registra los cambios de ciertos sitios web desde su creación.

The screenshot shows the Wayback Machine homepage. At the top, there's a navigation bar with links for WEB, BOOKS, VIDEO, AUDIO, SOFTWARE, and IMAGES. On the right side of the bar are links for SIGN UP | LOG IN, UPLOAD, and a search bar. Below the bar, there are links for ABOUT, BLOG, PROJECTS, HELP, DONATE, CONTACT, JOBS, VOLUNTEER, and PEOPLE. To the right of these are social media sharing icons for Facebook and Twitter, and a "Share on Facebook" button. The main content area features the "INTERNET ARCHIVE" logo and the "Wayback Machine" logo with a red "DONATE" button. A banner above the search bar says "Explore more than 866 billion web pages saved over time". Below the banner is a search input field with placeholder text "Enter a URL or words related to a site's home page". Underneath the search bar is a grid of 12 thumbnail images representing different web pages from various years. At the bottom of the page are four sections: "Tools" (with links to Availability API, Chrome Extension, Firefox Add-on, Safari Extension, MS Edge Add-on, iOS app, and Android app), "Subscription Service" (describing Archive-It), "Collection Search" (with a search form), and "Save Page Now" (with a URL input field and "SAVE PAGE" button).

INTERNET ARCHIVE

Wayback Machine

DONATE

Explore more than 866 billion web pages saved over time

Enter a URL or words related to a site's home page

Share on Facebook

Tools

Subscription Service

Collection Search

Save Page Now

Wayback Machine Availability API

Chrome Extension

Firefox Add-on

Safari Extension

MS Edge Add-on

iOS app

Android app

Archive-It enables you to capture, manage and search collections of digital content without any technical expertise or hosting facilities. Visit Archive-It to build and browse the collections.

Enter any keyword

End Of Term (US Gc)

SEARCH

This service is based on indexes of specific data from selected Collections.

https://

SAVE PAGE

Capture a web page as it appears now for use as a trusted citation in the future.

Problemas con los datos en la web actual

- Datos distribuidos
- Alto porcentaje de datos volátiles
- Grandes volúmenes
- Datos no estructurados y redundantes
- Calidad de los datos
- Datos heterogéneos



KEEP
CALM
AND
MAKE
BEAUTIFULL WEBSITES

Problema

Se desea crear un sistema para analizar la opinión con respecto a cierto tema. Las opiniones serán extraídas de varios sitios web y plataformas sociales; algunas se conocen con antelación pero el sistema debe de ser capaz de detectar otros sitios de interés para el análisis y tomar la información que interese procesar. La manera de recopilar la información tiene que ser eficiente.

¿Qué se puede hacer o usar para resolver el problema anterior?

Problema

Se desea crear un sistema para analizar la opinión con respecto a cierto tema. Las opiniones serán extraídas de varios sitios web y plataformas sociales; algunas se conocen con antelación pero el sistema debe de ser capaz de detectar otros sitios de interés para el análisis y tomar la información que interese procesar. La manera de recopilar la información tiene que ser eficiente.

¿Qué se puede hacer o usar para resolver el problema anterior?

Se necesita:

- un proceso para rastrear los sitios web de interés y,
- un proceso para obtener los datos de los sitios web.

Web Crawling

Proceso automatizado de navegar por la web de manera sistemática para indexar y recopilar información de diferentes sitios web.

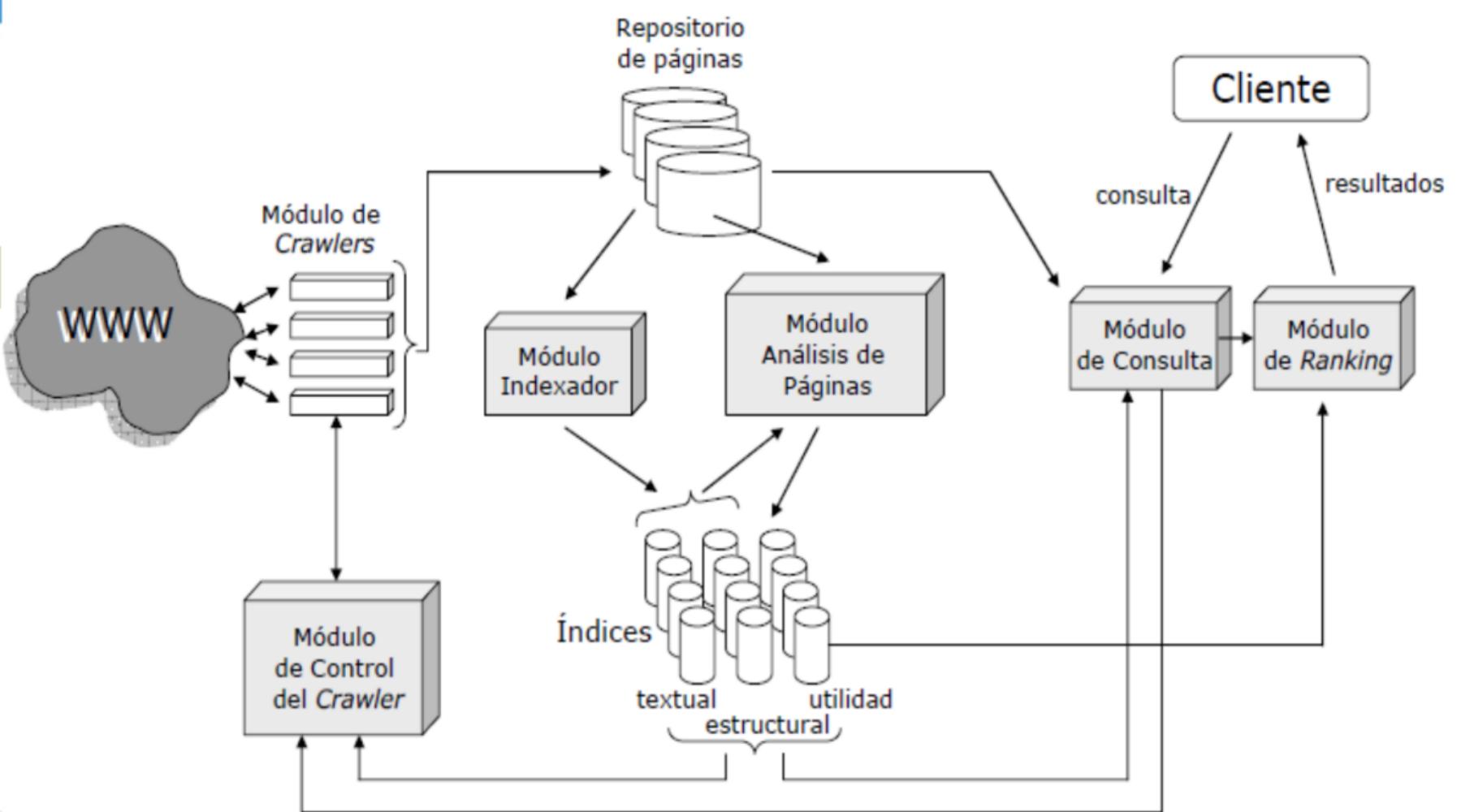
Tiene como objetivo:

- Recuperar, de manera eficiente y rápida, todos los recursos “importantes” de la web.

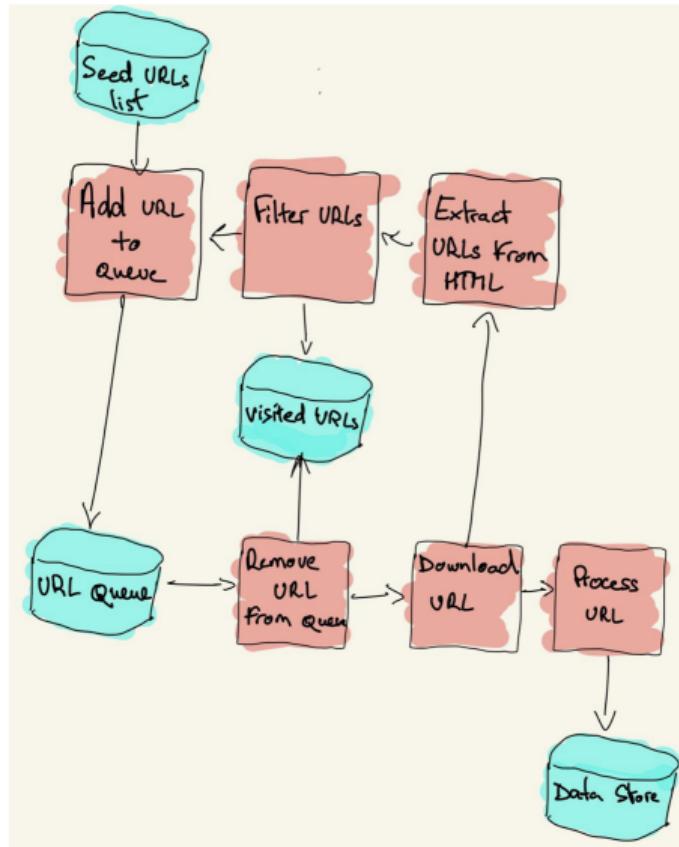
Los software que realizan la tarea de Web Crawling se les conocen como:

- crawler
- spider
- walker

Arquitectura general de un buscador web basado en crawlers

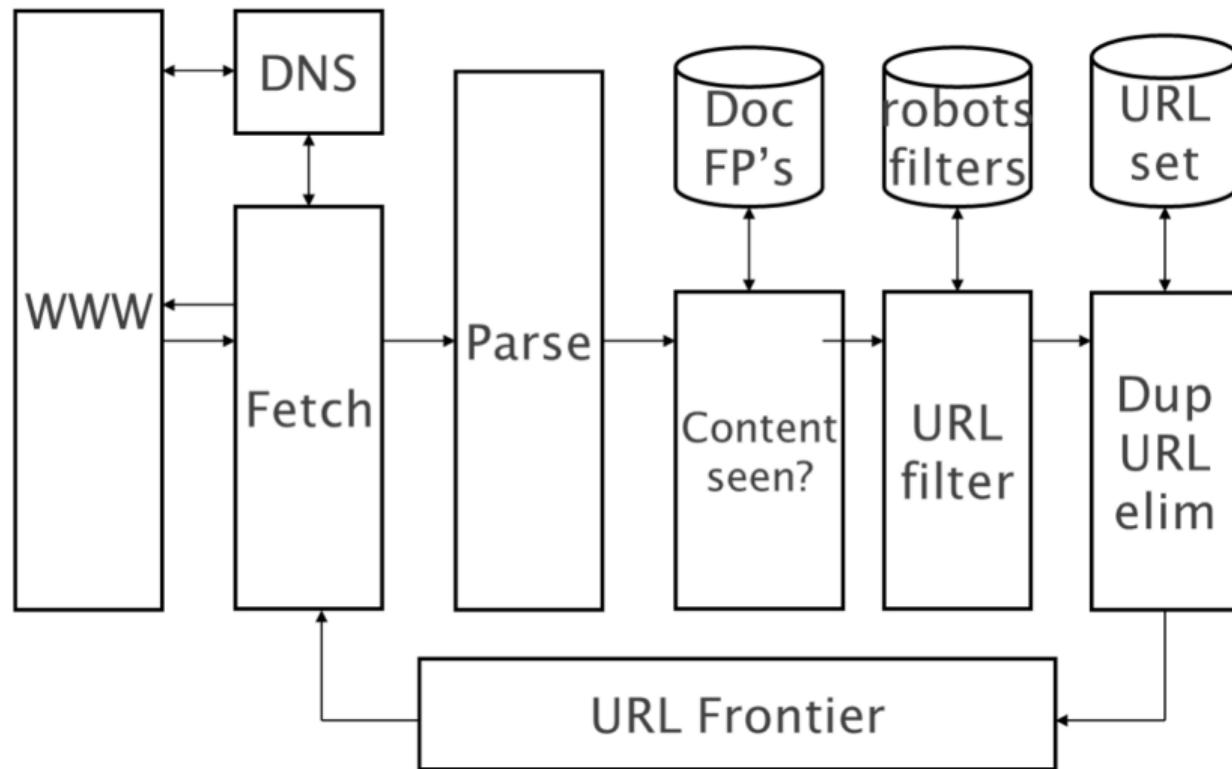


Procedimiento del crawler



Tomado de <https://www.scrapingbee.com/blog/crawling-python/>

Arquitectura y componentes formales del crawler



Políticas de los crawlers

Sobre el funcionamiento de los crawlers se definen ciertas reglas o políticas con el objetivo de

- Optimizar los recursos de red y,
- Evitar la saturación de los servidores.

Políticas usadas:

- Políticas de amabilidad
- Políticas de ordenación de URL
- Políticas de revisitado

Políticas de amabilidad

Las políticas de amabilidad, conocidas también como **políticas de rastreo**, son reglas establecidas en los sitios web para regular el comportamiento de los web crawlers.

Reglas:

- Frecuencia de rastreo
- Tiempo de espera entre solicitudes
- Exclusión de ciertas áreas del sitio
- Identificación del crawler

Políticas de ordenación de URL

Estas políticas consisten en un conjunto de reglas utilizadas para determinar el orden en el que se acceden a las URLs durante el proceso de rastreo y actúa directamente en la eficiencia del rastreo y en garantizar que los recursos del servidor se utilicen de manera efectiva.

Algunas de las reglas más utilizadas son:

- Aleatoriamente
- FIFO (First In, First Out)
- LIFO (Last In, First Out)
- Backlink Count
- Weighted Backlink Count
- Focused Crawling
- Batch PageRank
- PageRank

Políticas de revisitado

Estas políticas declaran reglas para definir cuándo y con qué frecuencia de visitan nuevamente las URLs.

- Uniforme
- Proporcional (computa previamente métricas)

Medidas para volver a rastrear una URL

- Edad
 - Se refiere al tiempo que ha transcurrido desde que fue descubierta la URL por primera vez por el crawler o desde su última visita.
 - Se calcula como:

$$Ep(url) = \begin{cases} 0 & \text{si la URL no ha sido modificada} \\ \text{fecha_actual} - \text{última_visita_a(url)} & \text{e.o.c.} \end{cases}$$

- Frescura
 - Se refiere a la frecuencia con la que se actualiza el contenido de la página y la importancia de las actualizaciones recientes.
 - Se calcula como:

$$Fp(url) = \begin{cases} 1 & \text{si el contenido de la URL es el mismo al almacenado} \\ 0 & \text{e.o.c.} \end{cases}$$

Recordando el problema

Se desea crear un sistema para analizar la opinión con respecto a cierto tema. Las opiniones serán extraídas de varios sitios web y plataformas sociales; algunas se conocen con antelación pero el sistema debe de ser capaz de detectar otros sitios de interés para el análisis y tomar la información que interese procesar. La manera de recopilar la información tiene que ser eficiente.

¿Qué se puede hacer o usar para resolver el problema anterior?

Se necesita:

- un proceso para rastrear los sitios web de interés y,
- un proceso para obtener los datos de los sitios web.

Web scraping

Proceso de recopilación de datos de páginas web de forma automatizada.

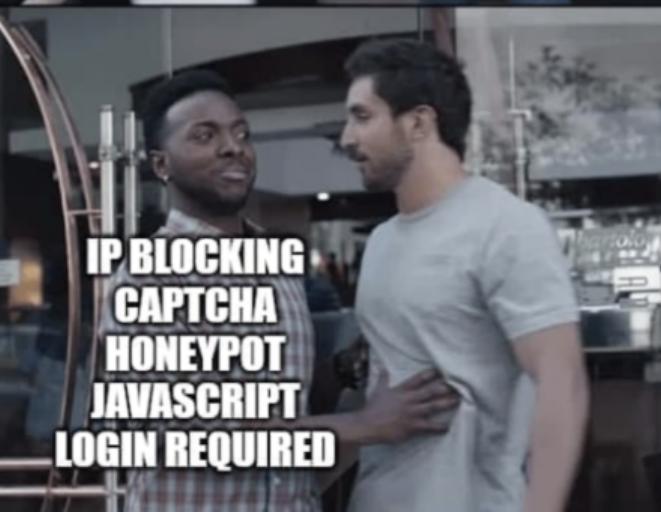
Utiliza programas de software para analizar y extraer información específica de sitios web, convirtiendo el contenido no estructurado en datos estructurados que pueden ser almacenados y analizados.

El web scraping también se conoce como **extracción de la web**.



WEB SCRAPING

DEVELOPERS



**IP BLOCKING
CAPTCHA
HONEYPOD
JAVASCRIPT
LOGIN REQUIRED**

Conclusiones

- La Web está en constante transformación y exige cambios de mentalidad en el diseño e implementación de los recursos que se comparten.
- La web constituye el mayor repositorio de información por lo que es imprescindible localizar, acceder y recopilar información que satisfaga las necesidades de un usuario.
- El crawler y el scraper son herramientas que posibilitan recuperar información en la web.

Bibliografía

- Oliva-Santos, R. (2008) Socialización de la información y el conocimiento en la Web. Revista Ciencias Matemáticas. Volumen 24 de 2008. Páginas 93-103.
- Baeza-Yates, R., Ribeiro-Net, B. (2002) Modern Information Retrieval. Páginas 367-395.
- Manning, Christopher D.; Raghavan, Prabhakar; Schütze, Hinrich (2007). An Introduction to Information Retrieval. Páginas 399-436.

Minería en la web: búsqueda, crawling y scraping

Sistemas de Recuperación de Información

Lic. Carlos León González
Dra.C. Lucina García Hernández

Facultad de Matemática y Computación
Universidad de La Habana

17 de junio de 2024