Ingeniería Informática

FACULTAD DE INGENIERIA
UNIVERSIDAD NACIONAL DE JUJUY

OPTATIVA

Recuperación Avanzada de Información

Dr. J. Federico Medrano
@jfedemedrano

Unidad N° 2 – Parte 1

Temas a desarrollar

- *Diseño de un spider/crawler. Estrategias.*
- *Web scraping.*
- *Recolección y representación estructural de sitios web/dominios.*
- *Tratamiento de grafos.*
- Minería de textos. Minería de datos.
- Recolección de repositorios. Protocolos de recuperación.
- Diseño de un motor de búsqueda. Diseño de un indexador de documentos.
- Optimización en motores de búsqueda web (SEO/SEM)

Diseño de un spider/crawler

Recolección de Información

- Cualquier estudio serio acerca de su estructura y funcionamiento debe partir de un número de páginas razonable.
- Algunos tipos de análisis requieren datos de dominios completos.
- No es posible hacer una recogida manual de datos.
- La información cuantitativa que ofrecen algunos servicios (buscadores, directorios) es poco fiable.
- Es necesario abordar una recogida automática de información

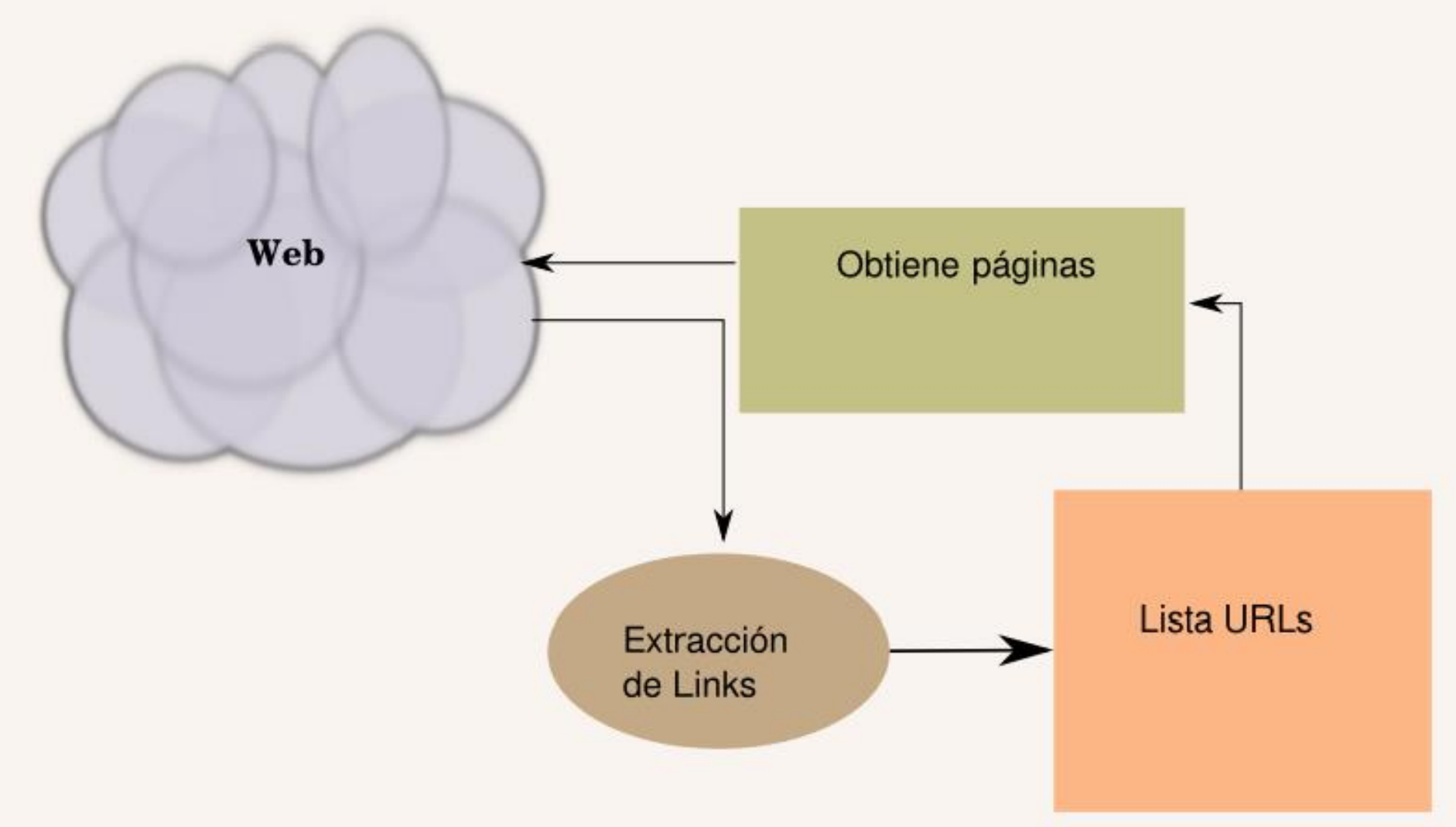
Qué es un crawler

Un **crawler** es un programa que navega de forma autónoma por la Web:

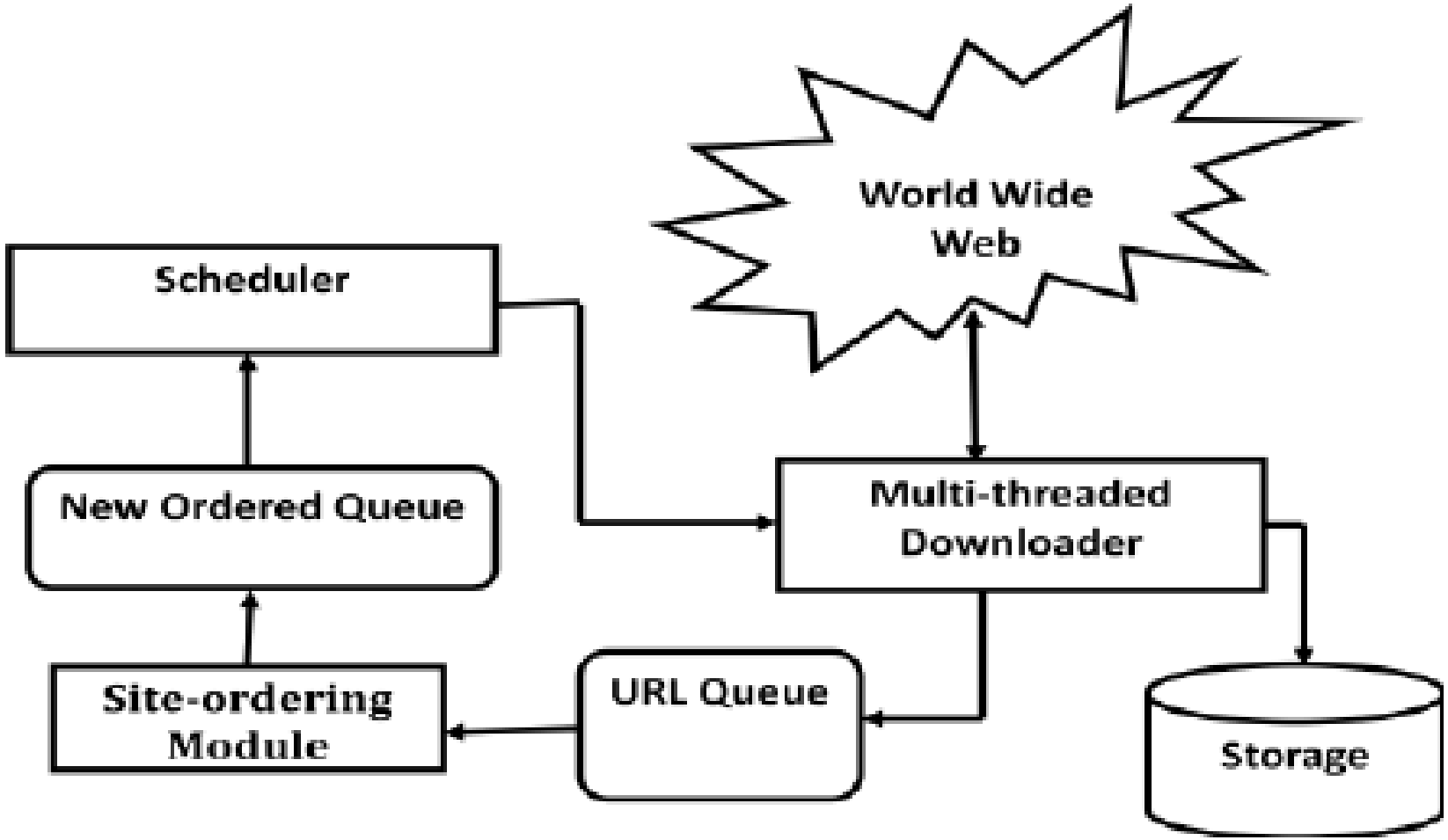
- Uno de sus usos es recoger información de cada página navegada para estudios *cibermétricos*.
- La mayor parte de los estudios *cibermétricos* necesitan utilizar un **crawler** para recoger datos.
- Los **crawlers** tienen también otros usos:
  - la detección de enlaces muertos
  - errores de codificacion
  - tareas de *mirroring* de sitios
  - la Recuperación de Información

Shiri (1998), el término ***Cybermetrics*** se refiere al análisis, estudio y medición cuantitativa de todas las clases y de todos los medios de información que existen en el Ciberespacio, usando para ello técnicas biblio-ciencio-informétricas.

Componentes básicos (1)



Componentes básicos (1)



La Web es difícil de explorar

- Su tamaño
- Su rapidez de cambio
- Las páginas dinámicas

Conclusión:

- No vale la fuerza bruta
- Hay que fijar estrategias de trabajo

Estrategias para:

- Selección de páginas a visitar
- Elección del camino u orden de visita
- Reglas de cortesía
- Optimizar la velocidad de exploración
- Políticas de refresco de páginas

Selección de páginas a visitar (1)

Por segmento web:

- En ocasiones se quiere explorar sólo un sitio web, o un dominio

Por tipo de recurso

- normalmente se quieren evitar recursos con ciertas extensiones: .gif, .jpg, .iso
- en algunos casos es normal limitarse a html/xml
- para Recuperación de Información se desearán explorar PDF, DOC,

Selección de páginas a visitar (2)

- Para evitar trampas para spiders

Una **spider trap** es una situación que hace que un crawler se encuentre en un círculo sin fin, explorando indefinidamente los mismos nodos

- Muchas no son intencionadas, pero otras sí:

Interesante: http://danzcontrib2.free.fr/en/pieges.php

- No hay manera de prevenir todas las trampas. La experiencia y las medidas ad-hoc son la solución más efectiva

Selección de páginas a visitar (3)

- Cuando un autor no quiere que su sitio sea copiado o indexado por los motores de búsqueda, puede usar:
  - Una *meta tag* como <meta name = "robots" content = "noindex, nofollow">.
  - Un archivo *robots.txt* que indica las partes del sitio que no se deben explorar.
  - .*htaccess* para prohibir los robots conocidos o detectados (cualquier webbot).

Trampas para crawlers

- URLs muy largos, que podrían producir un crash en el analizador léxico del crawler
- Direcciones recursivas
- Calendarios

Trampas para crawlers. *Soluciones*

- Eliminar de la exploración URLs que contengan determinadas subcadenas.
- Restringir exploración por longitud de URL
- Restringir por niveles de path
- Solución radical: no seguir URLs con "?"

Normalización de URLs

- Consiste en convertir las URLs a una forma estándar y consistente.
- El objetivo principal es detectar URLs iguales
- Algunos cambios son simples:
  - convertir a minúsculas
  - eliminar puerto por defecto
  - poner en mayúsculas las secuencias de escape

Normalización de URLs

Otros cambios tienen diversas implicaciones:

- Eliminar el fragmento o ancla:

http://www.httrack.com/html/abuse.html#WEBMASTERS = http://www.httrack.com/html/abuse.html

- Limitar protocolos:

ftp://ftp.rediris.es != http://ftp.rediris.es

- Ordenar variables (páginas dinámicas)

http://diaweb.usal.es/diaweb20/personal/presentacion.jsp?persona=30&tipo=P = http://diaweb.usal.es/diaweb20/personal/presentacion.jsp?tipo=P&persona=30

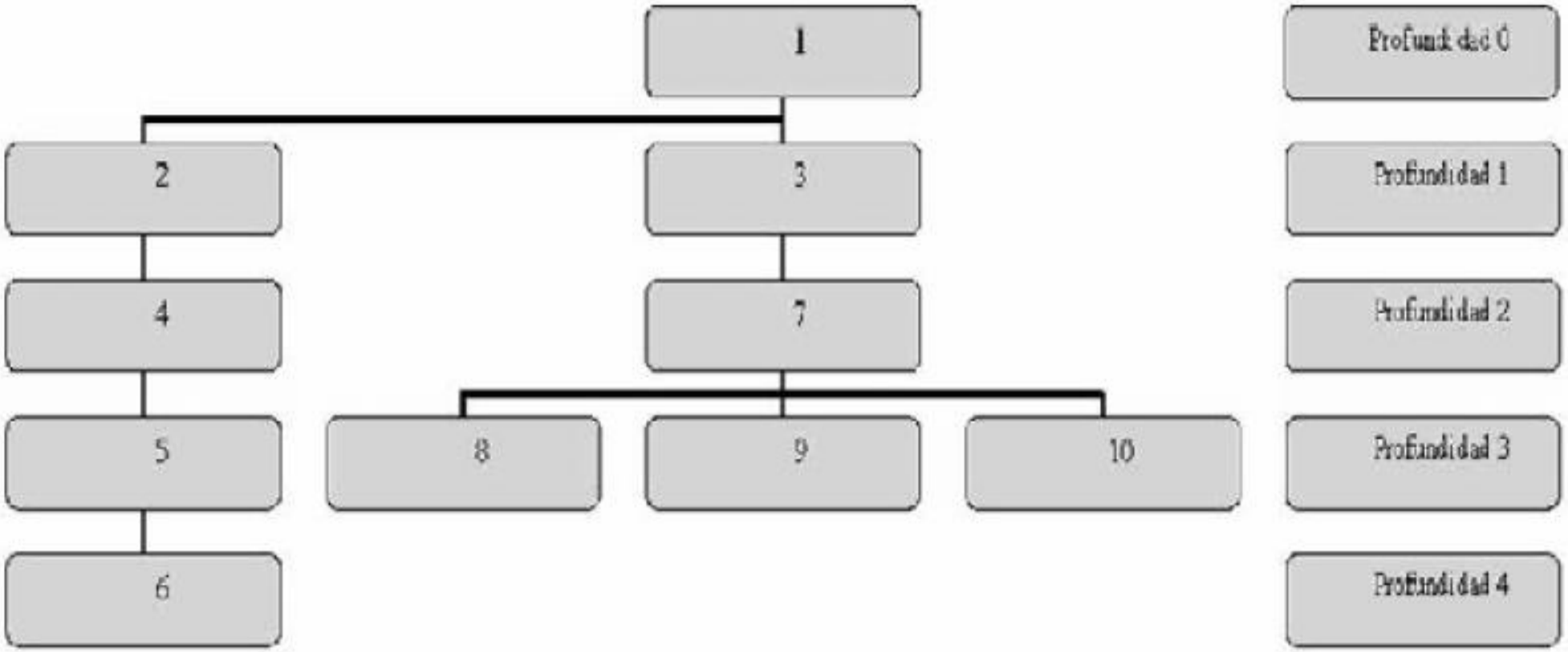
- Eliminar valores por defecto en variables

http://www.example.com/display?id=&sort=ascending = http://www.example.com/display

Elección del camino u orden de visita

- Se trata de priorizar la lista de URLs a explorar de la manera que más se adecúe a los objetivos del crawler.
  - en anchura
  - en profundidad
  - mejor candidato

Elección del camino u orden de visita



Recorrido en anchura: 1-2-3-4-7-5-8-9-10-6

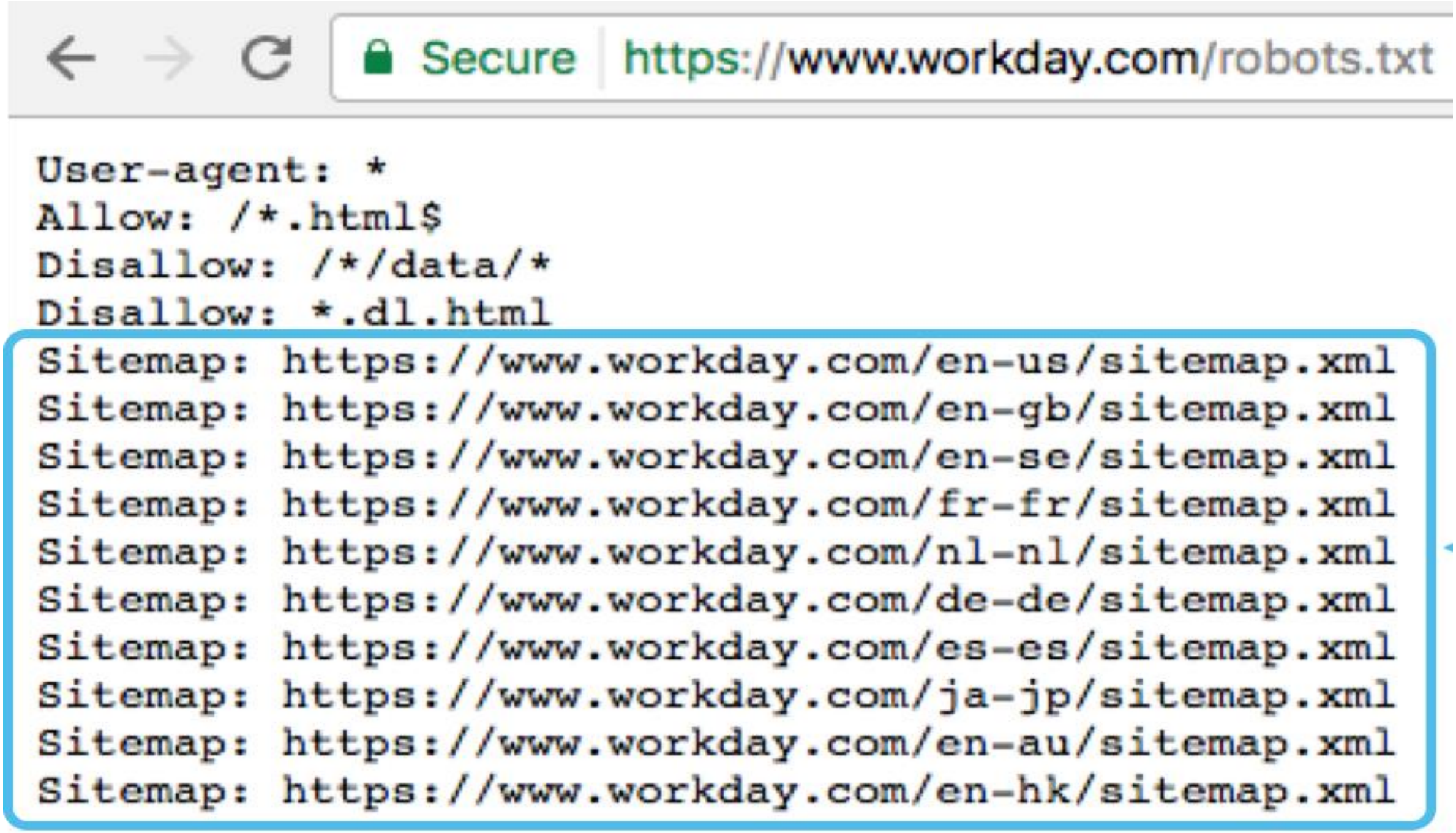
Recorrido en profundidad: 1-2-4-5-6-3-7-8-9-10

Reglas de cortesía

- Los crawlers consumen ancho de banda y recursos de los servidores
- Pueden ser vistos como amenazas por los administradores de sitios web
- Podrían ser bloqueados por los administradores
- Protocolo de Exclusión de Robots
- Intervalos entre visitas

Protocolo de Exclusión de Robots (SRE)

- Una convención que permite especificar:
  - quien puede explorar el sitio
  - que partes del sitio se pueden explorar
- Toma la forma de un fichero llamado robots.txt situado en el raíz de un servidor
- Es una convención que no 'obliga'. Los crawlers pueden saltarse el SRE


Secure
https://www.workday.com/robots.txt
User-agent: *
Allow: /*.html$
Disallow: /*/data/*
Disallow: *.dl.html
Sitemap: https://www.workday.com/en-us/sitemap.xml
Sitemap: https://www.workday.com/en-gb/sitemap.xml
Sitemap: https://www.workday.com/en-se/sitemap.xml
Sitemap: https://www.workday.com/fr-fr/sitemap.xml
Sitemap: https://www.workday.com/nl-nl/sitemap.xml
Sitemap: https://www.workday.com/de-de/sitemap.xml
Sitemap: https://www.workday.com/es-es/sitemap.xml
Sitemap: https://www.workday.com/ja-jp/sitemap.xml
Sitemap: https://www.workday.com/en-au/sitemap.xml
Sitemap: https://www.workday.com/en-hk/sitemap.xml

Secure https://www.buzzfeed.com/robots.txt

```
User-agent: msnbot
Crawl-delay: 120
Disallow: /*.xml$
Disallow: /buzz/*.xml$
Disallow: /category/*.xml$
Disallow: /mobile/
Disallow: *?s=mobile
Disallow: *?s=lightbox
Disallow: /bfmp/
Disallow: /buzzfeed/
Disallow: /contest
Disallow: /contests
Disallow: /plugin/
Disallow: /embed/
Disallow: /_comments/
```

Buzzfeed.com wants msnbot to wait 120 msc before crawling each page and NOT crawl any of these URL strings.

AND

```
User-agent: *
Disallow: /buzz/*.xml$
Disallow: /category/*.xml$
Disallow: /mobile/
Disallow: *?s=lightbox
Disallow: /bfmp/
Disallow: /buzzfeed/
Disallow: /contest
Disallow: /contests
Disallow: /_ga/
Disallow: /static/
Disallow: /dashboard/
Disallow: /plugin/
Disallow: /api/
Disallow: /buzzfeed/api/
Disallow: /embed/
Disallow: /_comments/
```

Buzzfeed.com wants all other user-agents (except for msnbot, discobot, and Slurp) to NOT crawl any of these URL strings

AND

```
User-agent: discobot
Disallow: /
```

Discobot should not crawl ANY URLs on buzzfeed.com.

AND

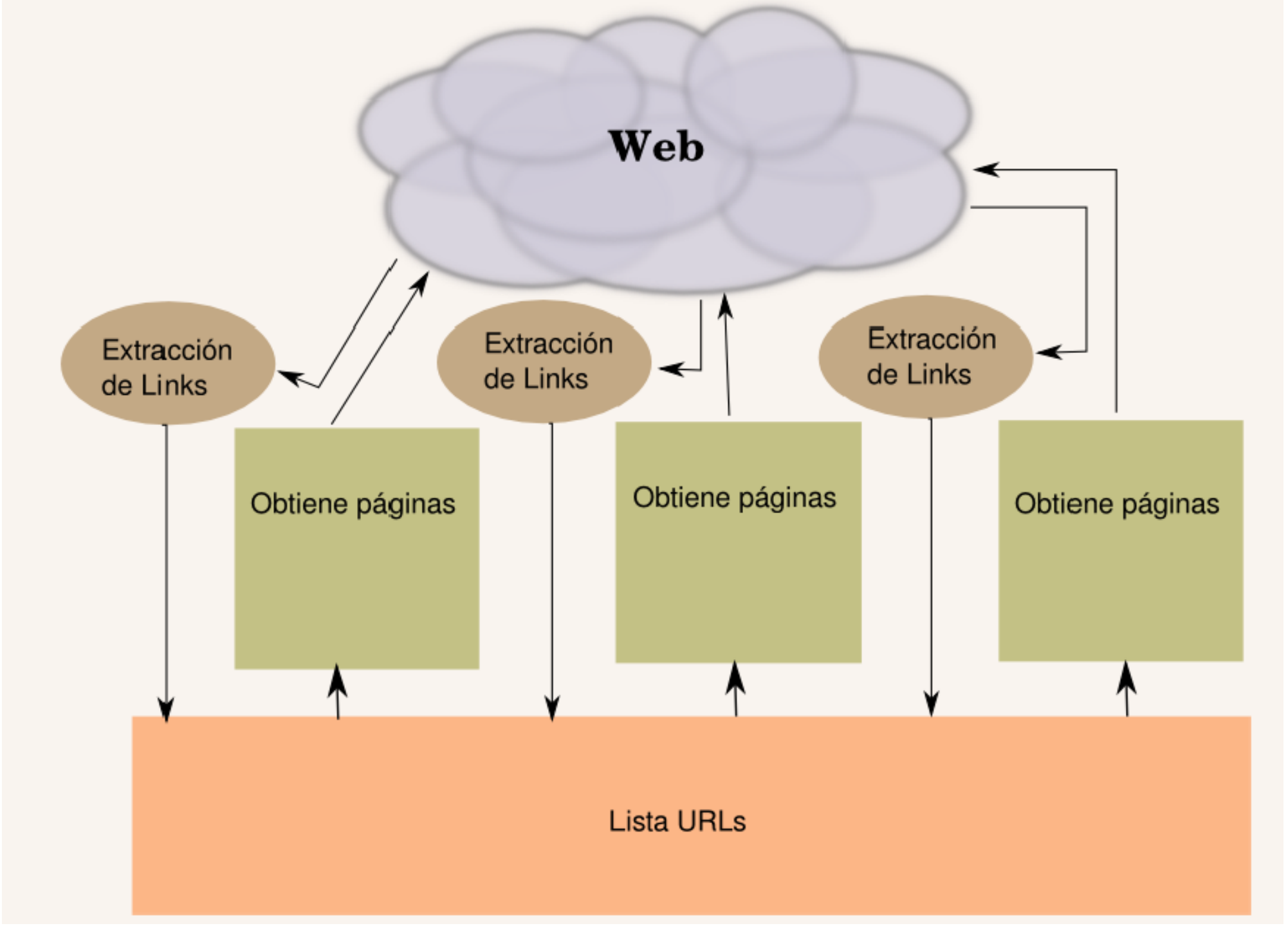
```
User-agent: Slurp
Crawl-delay: 4
```

Slurp (Yahoo's user-agent) should wait 4 msc before crawling each page, but crawl all URLs on buzzfeeed.com.

```
User-agent: *
Disallow: /buscadorWeb
Disallow: /buscador
Disallow: /alacarta/*.xml$
Disallow: /archivos/70-5561-FICHERO/www.tvemotogp%5B1%5D?download=1
Disallow: /archivos/70-6901-FICHERO/www.tvemotogp%5B1%5D
Disallow: /*.flv$
Disallow: /*.mp3$
Disallow: /*.mp4$
Disallow: /css/
Disallow: /rtve/components/parrilla/popup/
Disallow: /elecciones/css/i/*.jpg$
Disallow: /*.shtml?s1=
Disallow: /*?go=111b735a516af85c84df92fa7be3eface405339a91bac1c0d5bf8db96e9a2fc1
Disallow: /seriesmiticas/v/
Disallow: /sinatra/swf/flvplayer/
Disallow: /visor/flvplayer/
Disallow: /visordeportes/swf/flvplayer/
Disallow: /*entry.php?id=
Disallow: /comunes/publicidad/
Disallow: /contenidos/
Disallow: /concursocampusparty/files/
Disallow: /deportes/resultados/xml/
Disallow: /temporal/
Disallow: /noticias/temporal/
Disallow: /television/temporal/
Disallow: /radio/temporal/
Disallow: /deportes/temporal/
Disallow: /infantil/temporal/
Disallow: /*.inc$
Disallow: /su/
Disallow: /sm/
Disallow: /scdweb/
Disallow: /*SITE=es.antevenio.*
```

Paralelización

- Un crawler tiene tiempos de inactividad (espera de respuesta de servidores, etc.)
- Si incluye pausas o retardos entre peticiones estos tiempos de inactividad pueden ser mayores que los de actividad
- Pueden aprovecharse estos tiempos muertos para operar con otros servidores diferentes
- Un crawler multi-hilo permite aprovechar al máximo el ancho de banda disponible
- La velocidad de rastreo puede ser muy alta optimizando el aprovechamiento de esos tiempos muertos


Web
Extracción de Links
Extracción de Links
Extracción de Links
Obtiene páginas
Obtiene páginas
Obtiene páginas
Lista URLs

Web scraping

Web Scraping

- *Web scraping* es el proceso de recopilar información de forma automática de la Web.
- Usualmente, estos programas simulan la navegación de un humano en la Web ya sea utilizando el protocolo HTTP manualmente, o incrustando un navegador en una aplicación.
- El *web scraping* está muy relacionado con la indexación de la web, la cual indexa la información de la web utilizando un robot y es una técnica universal adoptada por la mayoría de los motores de búsqueda.

Web Scraping

- Sin embargo, el *web scraping* se enfoca más en la transformación de datos sin estructura en la web (como el formato HTML) en datos estructurados que pueden ser almacenados y analizados en una base de datos central, en una hoja de cálculo o en alguna otra fuente de almacenamiento.
- Alguno de los usos del *web scraping* son:
  - la comparación de precios en tiendas,
  - la monitorización de datos relacionados con el clima de cierta región,
  - la detección de cambios en sitios

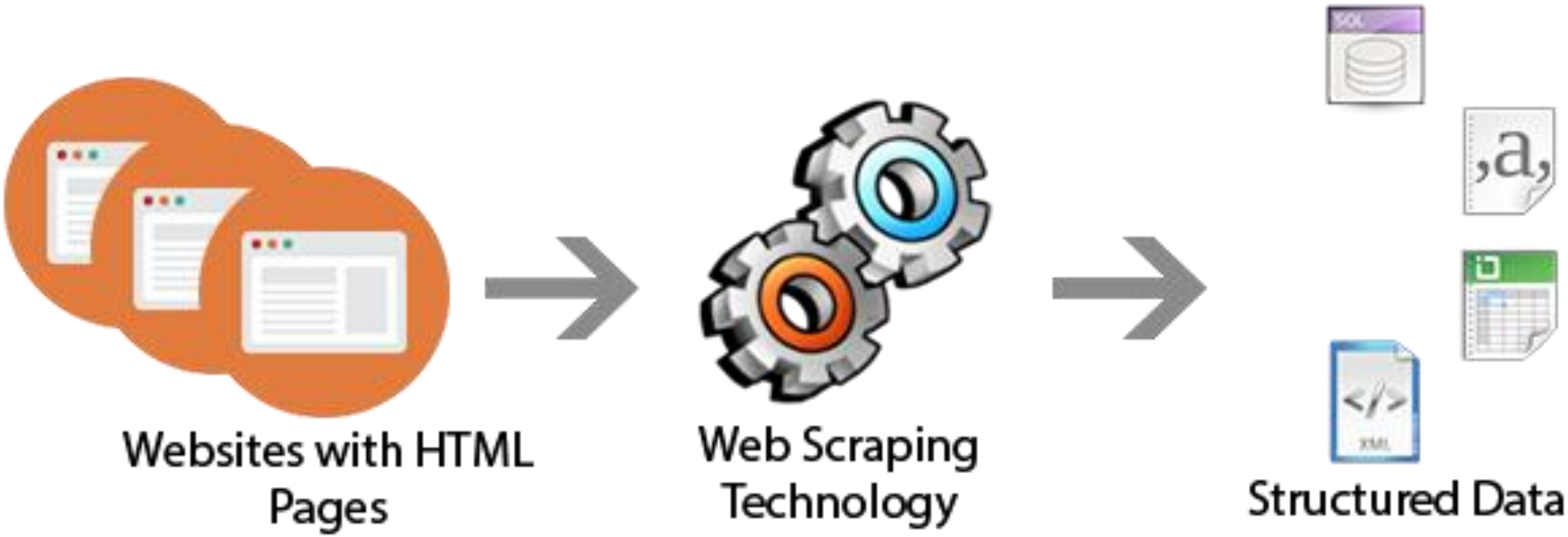
Web Scraping

- Su uso está muy claro: podemos aprovechar el **web scraping** para conseguir cantidades industriales de información (***Big data***) sin teclear una sola palabra. A través de los algoritmos de búsqueda podemos rastrear centenares de webs para extraer sólo aquella información que necesitamos.
- Para ello nos será muy útil dominar regex (regular expression) para delimitar las búsquedas o hacerlas más precisas y que el filtrado de la información sea mejor.

¿Cómo funciona un web scraper?

- Imaginemos que nos interesa extraer el título de 400 páginas que tienen el mismo formato y se encuentran dentro de un mismo site. En cada una de las 400 páginas el título está dentro de un selector <h1> que a su vez está dentro de un con la clase .header.
- Lo que hará nuestro web scraper es detectar ese selector h1 que está dentro de la clase header (.header h1) y extraerá esa información en cada una de estas 400 páginas. Luego podremos obtener toda esa información a través de la exportación de los datos en formatos como un listado en .json o un fichero .csv.

Web Scraping



Sitio web

- https://www.eltribuno.com/salta/nota/2020-4-21-20-43-0--somos-la-ultima-voz-en-primera-persona-para-contar-el-holocausto-dijo-una-sobreviviente

```
                            <div class="row air hidden-lg hidden-md hidde                                                    z-index:1;"><div class="publicidad box"> <div
data-spot=5d83ee1fa54a730013836fc3 data-id=5d83ef72a54a7300138371bb data-pos=P21 data-width='0' data-height='0' data-fixed='undefined' class='ppp ppp_P21'><div style=''>
<div class="gpt-load" data-gpt="/360145071/Notas_Mobile/Push_Notas_Mobile" data-size="[300, 250]"></div></div></div> </div></div></div>


                        <div class="category-wrapper">
                            <a class="category" href="#" itemprop="alternativeHeadline"><a href="/salta/ttag/helene-gutkowski" class="category" title="Helene Gutkowski"
alt="Helene Gutkowski" target="">Helene Gutkowski</a></a>
                                <span class="category pipe"> | </span>
                                <a href="/salta/personaje/polaco" class="category grouper">Polaco</a>
                        </div>

                        <div class="title-wrapper">
                            <h1 class="title newDetailTextChange" itemprop="headline">&amp;quot;Somos la última voz en primera persona para contar el Holocausto&amp;quot;, dijo
una sobreviviente</h1>
                        </div>

                        <p class="preview newDetailTextChange" data-link="2020-4-21-20-43-0--somos-la-ultima-voz-en-primera-persona-para-contar-el-holocausto-dijo-una-
sobreviviente">
                            <span class="detail-date">21 DE Abril 2020 - 21:07</span>
                            Helene Gutkowski es una de los miles de ni&amp;ntilde;os jud&amp;iacute;os que tuvieron que esconderse en familias cat&amp;oacute;licas en Francia para
salvarse del genocidio nazi.
                        </p>



                        <!--MINUTO A MINUTO-->

                            <div class="object-list-highlights air">
                                <ul class="highlights-wrapper">
                                        <li class="highlight-item newDetailTextChange">
                                            <i class="fa fa-play-circle fa-1" aria-hidden="true"></i>
                                            <a href="/salta/nota/2020-4-21-17-23-0-por-que-se-conmemora-hoy-el-dia-del-recuerdo-del-holocausto-y-el-heroismo"
class="highlight-link" target="">Por qué se conmemora hoy el Día del Recuerdo del Holocausto y el Heroísmo</a>
                                        </li>
                                        <li class="highlight-item newDetailTextChange">
                                            <i class="fa fa-play-circle fa-1" aria-hidden="true"></i>
                                            <a href="/salta/nota/2020-1-25-15-24-0-el-lunes-se-cumplen-75-anos-de-la-liberacion-de-auschwitz-por-el-ejercito-rojo"
class="highlight-link" target="">El lunes se cumplen 75 años de la liberación de Auschwitz por el Ejército Rojo</a>
                                        </li>
                                </ul>
                            </div>

                    </div>

                    <div class="row hidden-print">
                        <div class="col-xs-12">
                            <div class="detail-social-icons social-sharing-box sticky" data-hide-on-scroll=".extras.object-te-puede-interesar" data-social="tosticky">
                                <ul class="social-wrapper clearfix">
                                    <li class="social-icon fb no-shares">
                                        <a class="social-link social-facebook" onclick="javascript:var popup = window.open('https://www.facebook.com/sharer/sharer.php?
```

¿Problemas?

- El ***raspado*** no es del todo legal. Lo cual puede traer inconvenientes al recolector y al servidor (rendimiento).
- Los cambios frecuentes de estructura de un sitio web dificultan la extracción.
- .htaccess es un archivo de configuración para tu servidor web. Y **se puede modificar para evitar que los raspadores accedan a tus datos**.
- Recolecciones periódicas y mismas recomendaciones que para el diseño de crawlers.

Recolección y representación estructural de sitios web/dominios

La Web como un grafo

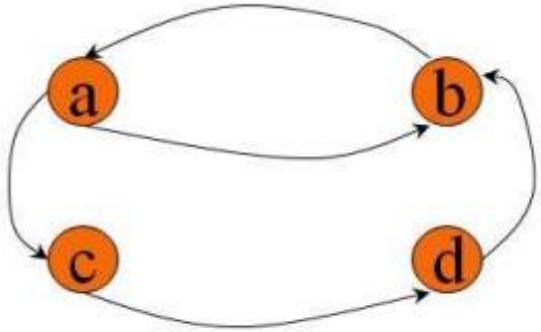
- La web puede verse como un enorme grafo dirigido.
- Las porciones pequeñas de la web (un sitio web, un dominio) siguen el mismo principio.
- Cada página web es un nodo, y un link (enlace) entre dos páginas representa una relación entre ambas (relación de entrada o salida)

Grafos

- Un grafo es un pareja G = (V, E), donde V es un conjunto finito de puntos llamados vértices o nodos y E es un conjunto de pares de puntos llamados aristas o arcos que conectan a los nodos de un grafo (E={x,y}; x, y ϵ V)
- V(G) = Es el conjunto de vértices.
- E(G) = Es el conjunto de conexiones del grafo.

Representación matricial

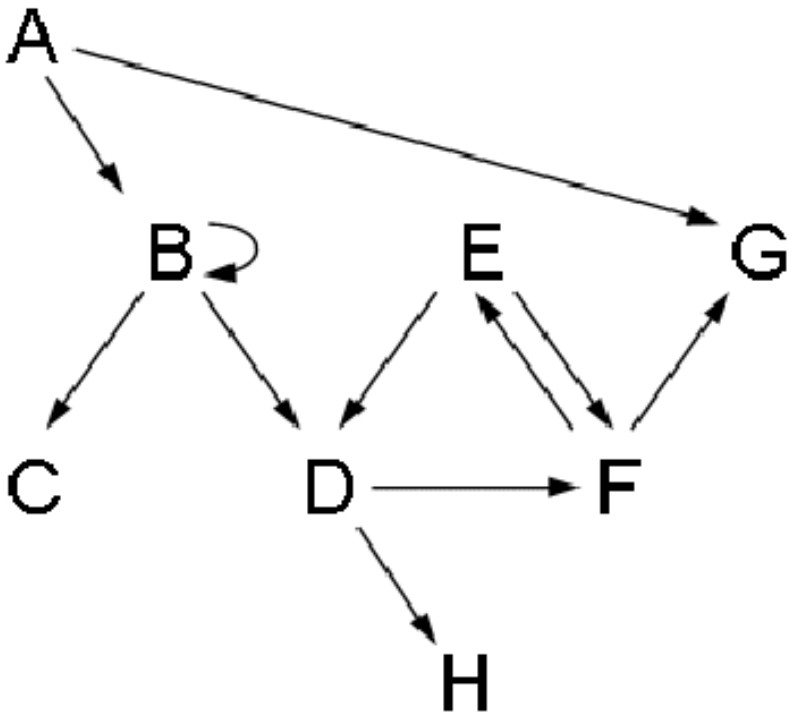
- Booleana
- Orden arbitrario a los vértices
- Filas y columnas el mismo orden
- Ventaja: tiempo de acceso
- Desventaja: espacio de almacenamiento
- Se puede determinar si existe un camino entre dos nodos



	a	b	c	d
a	0	1	1	0
b	1	0	0	0
c	0	0	0	1
d	0	1	0	0

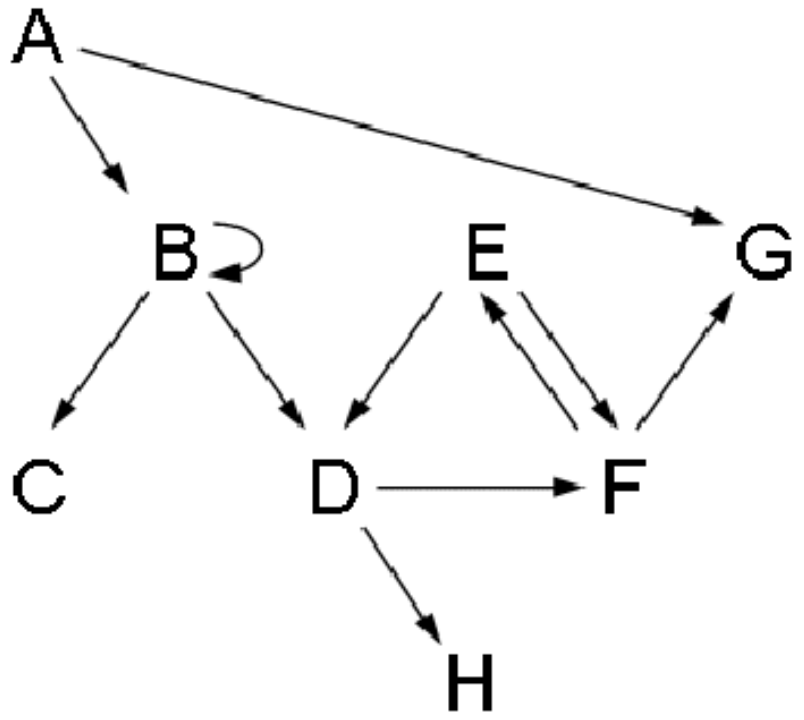
Análisis en Cibermetría

- B tiene un enlace entrante (inlink) desde A - citación.
- B tiene un enlace saliente (outlink) hacia C - referencia.
- B tiene un autoenlace (selflink) - autocita.
- A no tiene inlinks, no-linked – no citado
- D tiene un enlace interno (internal link) a F.



Análisis en Cibermetría

- A cita transitivamente a C y D a través de B.
- E y F están recíprocamente enlazados.
- A tiene un (hotlink, link transversal) enlace directo a G.
- B y E co-enlazan (co-linking) a D – documentos relacionados.
- C y D son co-enlazados (co-linked) por B - cocita.
- *El mismo enfoque para una web que para un artículo científico.*



Tipologías de páginas web

- Páginas que tienen muchos links que apuntan a ellas: autoridades.
- Páginas que tienen muchos links de salidas: conectores, hubs o índices.
- Regiones de la Web unificadas temáticamente muestran las mismas características que la Web en general.