Crawlers - Arañas

UCR – ECCI

CI-2414 Recuperación de Información

Prof. Kryscia Daviana Ramírez Benavides

¿Qué es una Araña?

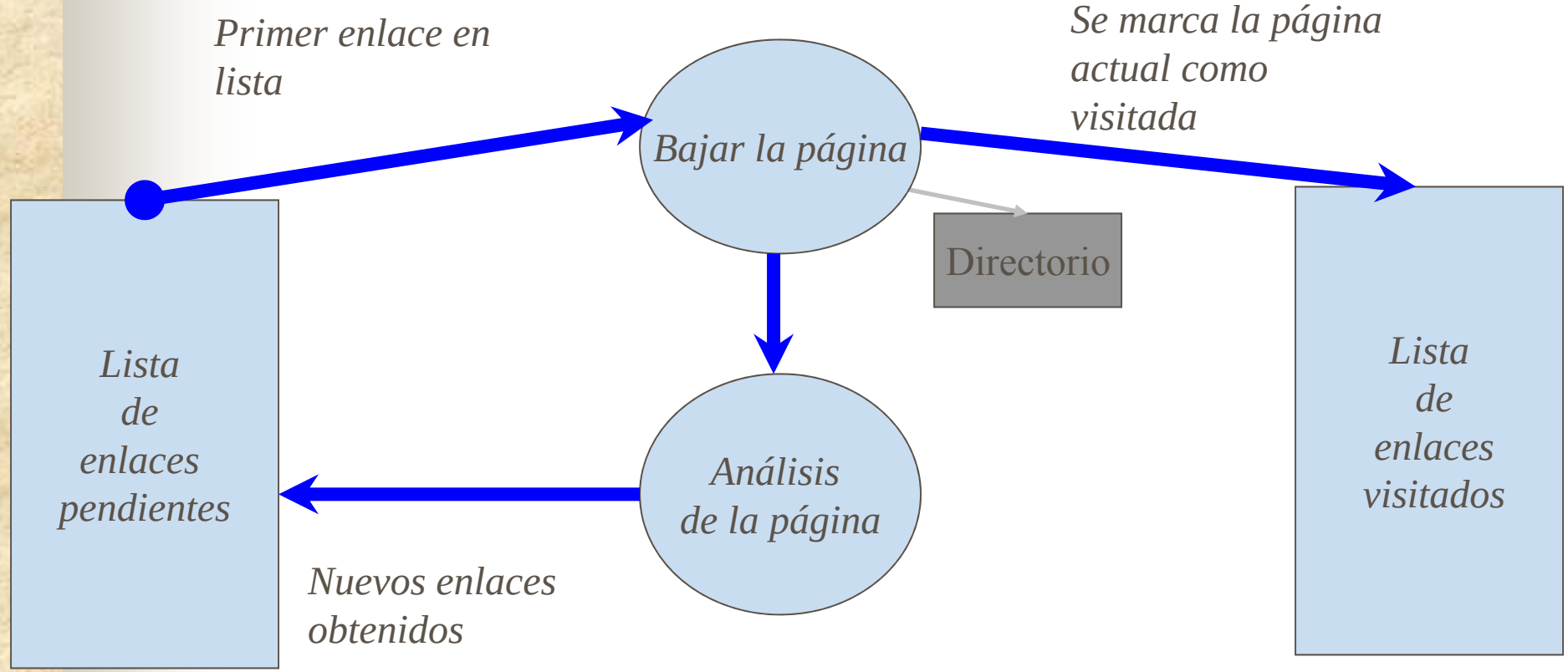
- También se le llama *robot* o *araña* (*spider*, *crawler*).
- Una araña (*crawler*) es aquella aplicación que es capaz de moverse a través de distintos sitios de la red.
- De manera que localiza los enlaces contenidos dentro de cada página para recuperar otras direcciones, ya sean "*links internos*, *externos*" u otros.
- Se ejecuta en una máquina local y envía peticiones a los servidores Web, visitas periódicas. Permite que se le proporcionen direcciones de sitios Web.

¿Qué es una Araña? (cont.)

- La función central de un *crawler* es revisar muchas páginas al mismo tiempo, para solapar los retrasos en:
  - Resolver los nombres de un URL a direcciones IP utilizando el DNS.
  - Conectar un *socket* al servidor y enviar la solicitud.
  - Recibir la página solicitada en respuesta.

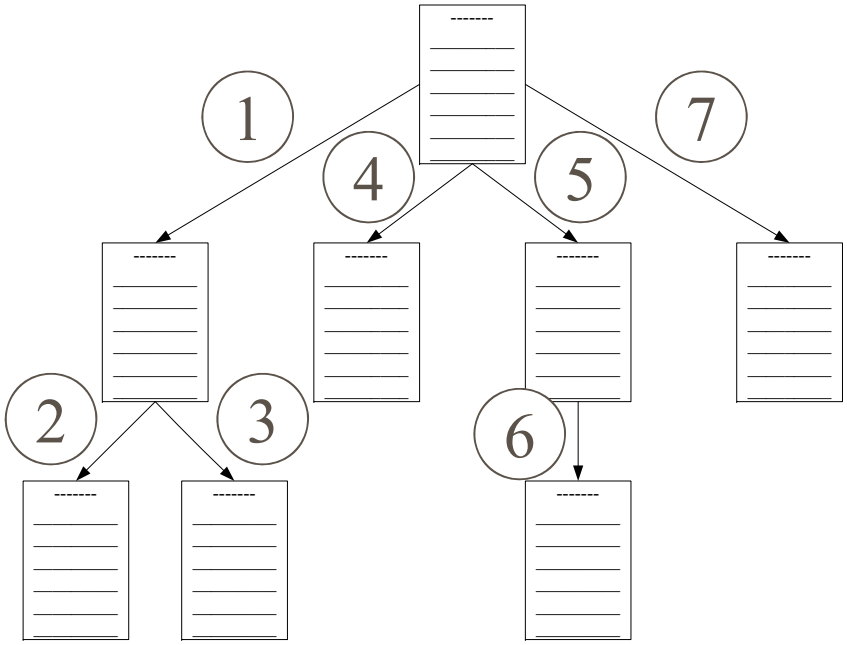
  junto con el tiempo que se gasta al escanear las páginas y salvar las páginas en un repositorio local.

Forma Básica de una Araña



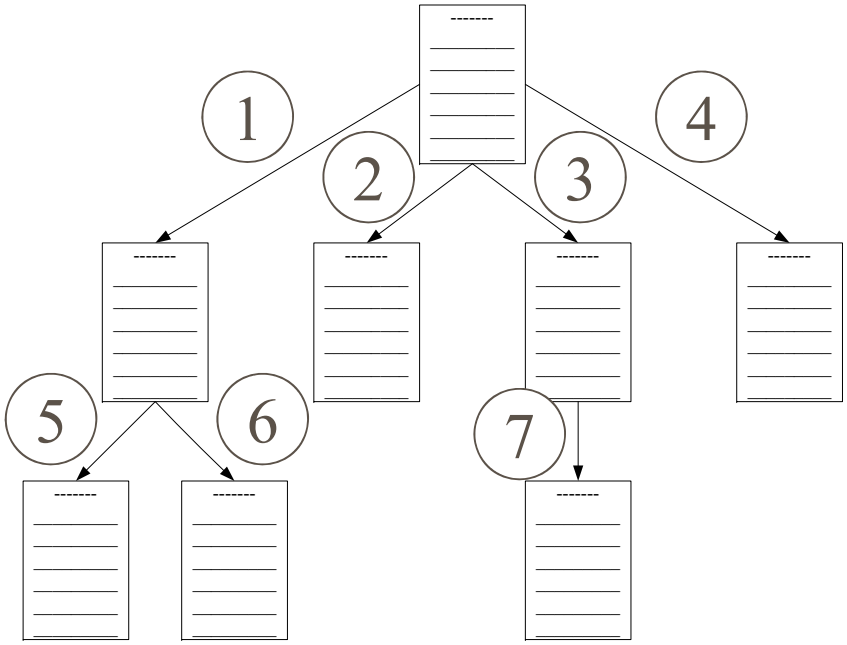
Tipos de Recorrido de las Arañas

- *Recorrido en profundidad:* Se sigue el primer enlace de una cadena de páginas hasta llegar a la más profunda, desde la que se vuelve recursivamente.



Tipos de Recorrido de las Arañas (cont.)

- *Recorrido en anchura:* Se examinan todas las páginas a las que se llega desde la página actual, visitándose después las del siguiente nivel.



Otras Técnicas

- Los buscadores combinan las técnicas anteriores con medidas de **popularidad** para decidir el orden en que se visitan las páginas.
- El objetivo es recorrer las páginas de **mayor calidad**.
- Popularidad se mide como el número de enlaces que apuntan a la página.
- Límite en la profundidad por sitio Web es el límite al número de páginas que puede devolvernos el buscador.
- **Frecuencia** entre visitas es **variable** (días – meses).

Limitaciones

- En el intervalo entre actualizaciones los buscadores pueden devolver *enlaces inválidos* (porcentaje de 2-9%).
- Los administradores de sitios Web pueden controlar el comportamiento de los *crawlers*, por ejemplo: impedirles que indexen determinadas páginas.
- No se indexan páginas generadas dinámicamente o protegidas con contraseñas.
- Los *crawlers* pueden tener *problemas* para indexar páginas con *frames* o mapas de imágenes.

Funcionamiento de una Araña

- Comienza con un conjunto de *URLs*, ya sea haciendo un recorrido en **profundidad** o **anchura**.
- Un *crawler* recupera un documento y recursivamente todos los documentos con los que mantiene vínculos dicha página, indexa la información de acuerdo a un criterio predefinido.
- Los criterios son: el título del documento, los metadatos, el número de veces que se repite una palabra en un documento, algoritmos para valorar la relevancia del documento, etc.; y el peso de cada criterio varía de acuerdo al motor de búsqueda.

Funcionamiento de una Araña (cont.)

- La información se almacena en una base de datos, la cual puede ser consultada por los usuarios de Internet para recuperar la información deseada.
- Para mantener actualizada la base de datos, los *crawlers* vuelven a visitar los sitios para verificar que las páginas registradas se mantengan activas, de no ser así (cuando se mueven a otro sitio o desaparecen) las eliminan de la base de datos.

Funcionamiento de una Araña (cont.)

- El *crawler* (robot o *spider*) recorre la red de forma automática explorando los servidores a nivel mundial, o en el ámbito de especialización del buscador (geográfico, idiomático o temático).
- Algunos *crawlers* se la pasan vagando y recolectando información para análisis estadísticos, otros ejecutan actividades de reflejo sobre un sitio (*mirroring*) con el fin de evitar sobrecargarlo (por ejemplo para bajar archivos de software), distribuyendo la carga en diferentes servidores de Internet. Otra de las razones es evitar que el cliente descargue la información de un servidor que se encuentre a una mayor distancia.

Funcionamiento de una Araña (cont.)

- Los *crawlers* ejecutan una tarea muy útil, pero consumen gran parte del ancho de banda, lo cual puede ser frustrante.
- Si no está bien programado puede crear ataques de servicio prohibidos y no deseados en algunos servidores, al intentar obtener la información a una velocidad mayor que la que soporta el servidor.

Funcionamiento de una Araña (cont.)

- La interacción humana también puede desconfigurar un *crawler*, o no entender el impacto que causará su configuración en los servidores que contacta.
- El mayor problema es la falta de inteligencia, es decir, ¿con base en qué toma la decisión el *crawler* para visitar paginas?
- A pesar de todos los problemas los *crawlers* ofrecen un servicio valuable a la comunidad de Internet.

Estructura de los Sitios Web

- La araña deberá conocer la estructura de la página en la que se encuentra.
- Esto lo hace examinando el código HTML en busca de las referencias (hipervínculos) a otras direcciones.
- Por ejemplo:

  Click Here

  Pese la variedad de instrucciones e información, la araña busca solo el atributo *href* e ignora el resto.

Estructura de los Sitios Web (cont.)

- Puede hallar tres tipos de enlaces:
  - Internos.
  - Externos.
  - Otros.

Estructura de los Sitios Web (cont.)

- ***Enlaces internos:*** Realiza una conexión a otra página que se encuentra en el mismo sitio en la red.
- Por ejemplo: *http://www.kimmswick.com/index.shtml* hace referencia a la dirección *http://www.kimmswick.com/attractions.shtml*.

Estructura de los Sitios Web (cont.)

- ***Enlaces externos:*** Realiza una conexión a una dirección externa al sitio en el que se está actualmente.
- Por ejemplo: Estar en *http://www.kimmswick.com/attractions.shtml* e ir a *http://www.yahoo.com*.

Estructura de los Sitios Web (cont.)

- ***Otros enlaces:*** No todos los enlaces deben seguir un protocolo *http*; por ejemplo está el siguiente tipo de enlace: *mailto:webmaster@kimmswick.com*.

Algoritmos de Crawling

- Se comienza desde un conjunto de URLs muy populares o enviados explícitamente por administradores de sitios Web, y se siguen los links desde allí recursivamente, evitando repeticiones.
- El recorrido puede ser una cobertura amplia (anchura) o cobertura vertical (profundidad).
- Es complicado coordinar varios *crawlers* para que no repitan trabajo. Una alternativa utilizada es que se repartan los dominios.

Algoritmos de Crawling (cont.)

- Lo que se ve en un índice es como las estrellas del cielo: jamás existió. Cada página se indexó en un momento distinto del tiempo (pero al ir a ella obtenemos el contenido actual).
- Las páginas suelen tener entre 1 día y 2 meses de antigüedad, 2%-9% de los links almacenados son inválidos.
- Las páginas enviadas a indexar explícitamente demoran pocos días o semanas en ser indexadas, las no enviadas entre semanas y dos meses.

Algoritmos de Crawling (cont.)

- Los mejores *crawlers* recorren unas 10 millones de páginas por día.
- Requerirían un mes para rotar todo el Web, en el mejor caso. En la práctica ni siquiera alcanzan a indexar gran parte.
- Para complicar las cosas, existen protocolos de buen comportamiento para *crawlers*, de modo que no saturen a los servidores Web.

Algoritmos de Crawling (cont.)

- Algunos *crawlers* aprenden la frecuencia de actualización de las páginas e indexan las más volátiles más frecuentemente, por ejemplo: la página de la CNN versus mi página personal.
- Google utiliza el algoritmo de ranking de páginas para determinar el orden de *crawling*, y obtiene muy buenos resultados.
- Las páginas generadas dinámicamente son actualmente uno de los problemas mayores para los *crawlers*.

Estructura de la Araña

- Hay dos maneras de crear una araña:
  - Por recursión.
  - Sin usar recursión.
- Para definir cuál esquema utilizar, se recomienda pensar en cuán larga será la profundidad de sitios Web a visitar.

Estructura de la Araña - Construcción Recursiva

- Basado en el llamado a sí mismo cada vez que se encuentra un enlace o hipervínculo.
- Uso de una sola pila de almacenamiento temporal.
- Es recomendable sólo cuando hay pocas páginas por visitar.
- No facilita el uso de la multiprogramación, ya que la recursión provocaría que cada hilo tuviese su propia pila.

Estructura de la Araña - Construcción Recursiva (cont.)

- Podría crear un ciclo infinito de llamadas y visitas a nuevas direcciones: al encontrar un enlace lo sigue inmediatamente sin terminar de examinar una página en ese momento.
- Si se da un error, se desaprovechará gran cantidad de enlaces no explorados en las páginas iniciales.

Estructura de la Araña - Construcción No Recursiva

- Supera los problemas de la recursividad.
- Poco más compleja de diseñar e implementar.
- A diferencia de la recursiva, esta estructura evalúa toda una página al mismo tiempo.
- Uso de colas:
  - Cola de espera.
  - Cola de ejecución.
  - Cola de errores.
  - Cola de URLs completados.

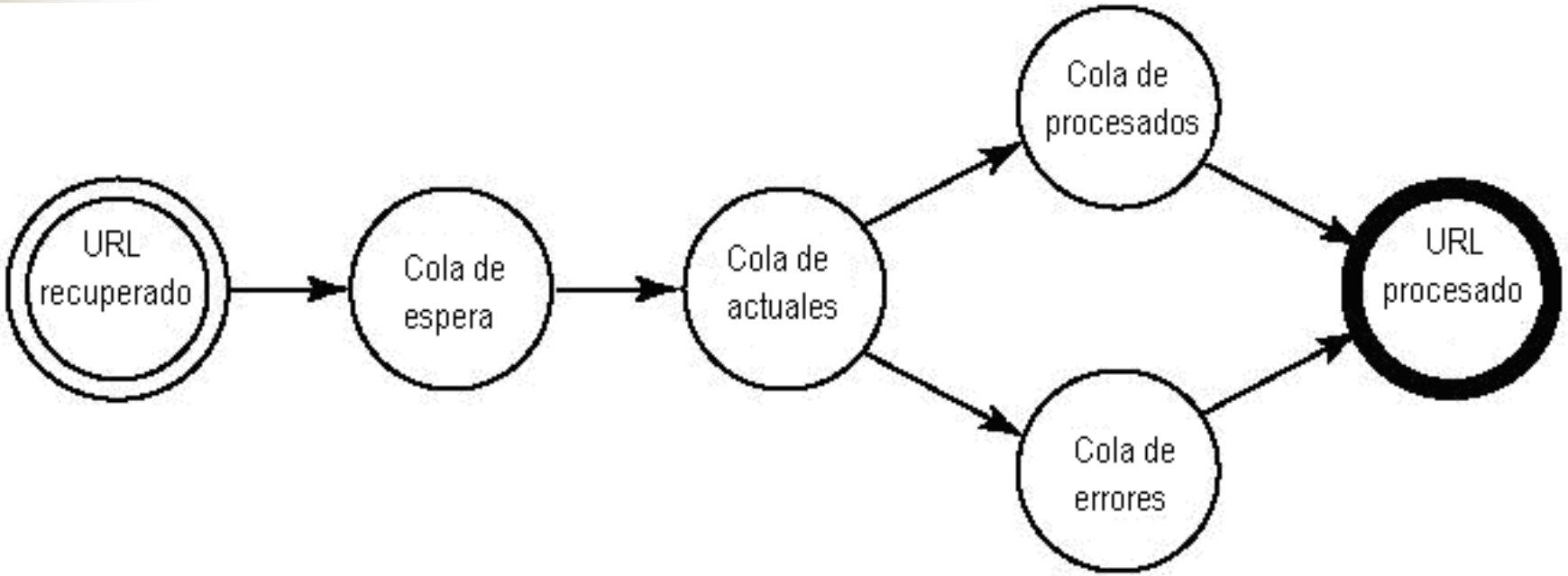
Ejemplo Sencillo de una Araña

- Originalmente esta implementación permite bajar todas las páginas relacionadas con un sitio específico.
- La araña inicia su procesamiento a partir de una dirección URL especificada por el usuario.
- Por cada página que analiza:
  - Pone cada dirección en una cola de espera.
  - Posterior a esto, al atender dicha página la saca de la cola de espera, la pone en la cola de atendidos, la baja y la guarda a una determinada dirección local.
  - Realiza un escaneo en dicha página, buscando más direcciones. Si encuentra más direcciones las pone en la cola de espera.
  - Una vez que la página ha terminado de ser examinada, se coloca en la cola de terminados para que quede en los respectivos registros de su visita.

Ejemplo Sencillo de una Araña (cont.)

- Las colas son almacenadas y utilizadas desde memoria RAM. Existe la posibilidad de utilizar una base de datos SQL para su almacenamiento.
- Utiliza una colección de hilos (*thread pool*) con un máximo de 100 hilos, que puede variar.
- Se debe especificar, además, el tamaño máximo de una página (en kilobytes) que es permitido bajar.

Estructura de la Araña



Problemas

- Volatibilidad
  - Crecimiento del 200% anual.
  - 23% cambian constantemente.
  - Promedio de vida de 10 días.
- Eficacia
  - 40% del Tráfico
- Espacio disponible
- Datos incompletos
  - Indexado solo el 12%.

Problemas (cont.)

- Tiempo
  - Visitas innecesarias.
  - Manejo de tiempo entre visitas a páginas.
  - Calendarización de la Araña.
  - Dificultad para establecer el periodo de actualización correcto.
- RI en el “Web Oculto”
  - Como obtener toda la información necesaria.
- Las páginas dinámicas.

Mejoras Sugeridas

- Recuperación de trabajo anterior.
- Distribución de Trabajo.
- Actualización de páginas de manera probabilística.
- Uso de *PageRanking*.
- Manejo de cambio de ubicación y muerte de páginas.
- Limitar alcance de la araña para personalización.
- Monitoreo (visualización de los cambios en el Web).

Referencias Bibliográficas

- La información fue tomada de:
  - Presentación realizada por Didier Cerdas, Curso Recuperación de Información, 2002.
  - Presentación realizada por Adriana Blanco, Francisco García y Daniel Mora, Curso Recuperación de Información, 2004.
  - Libro de texto del curso.