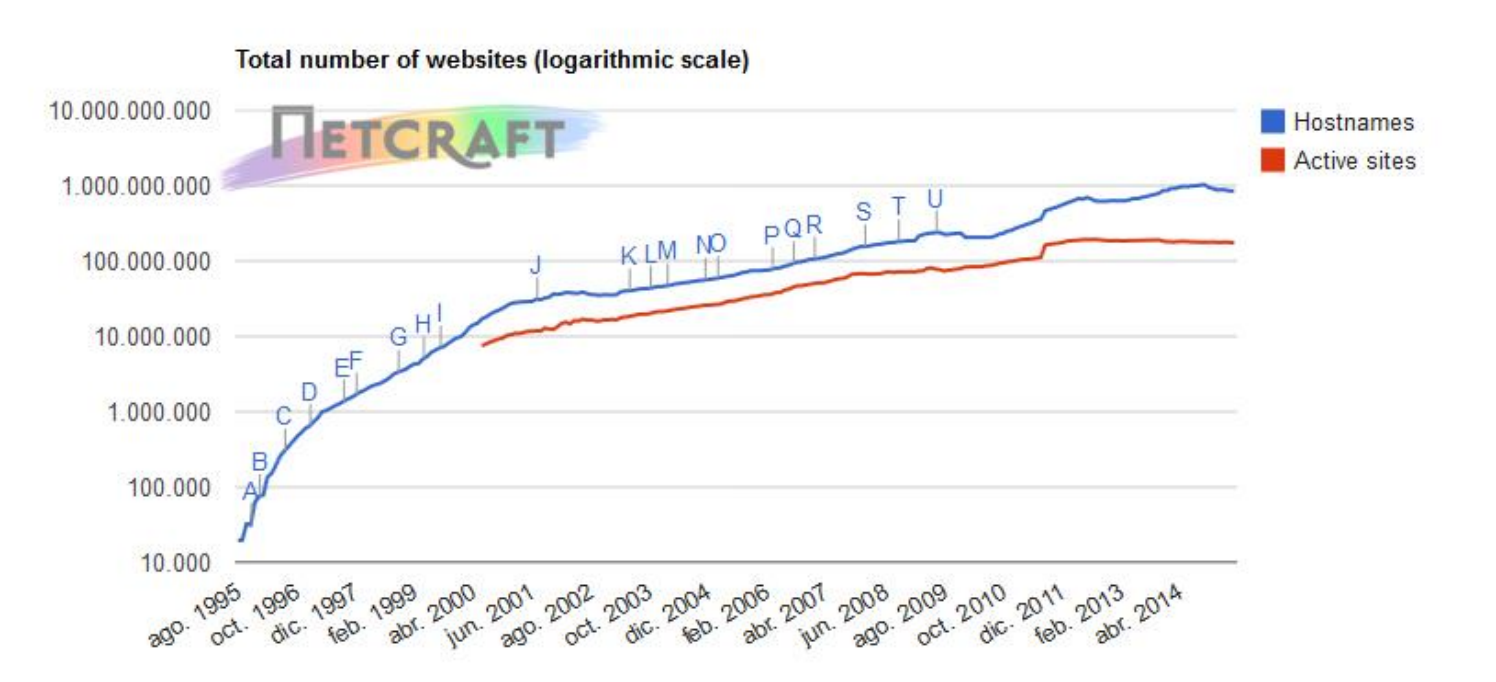
Ingeniería de Aplicaciones Web

Diego C. Martínez

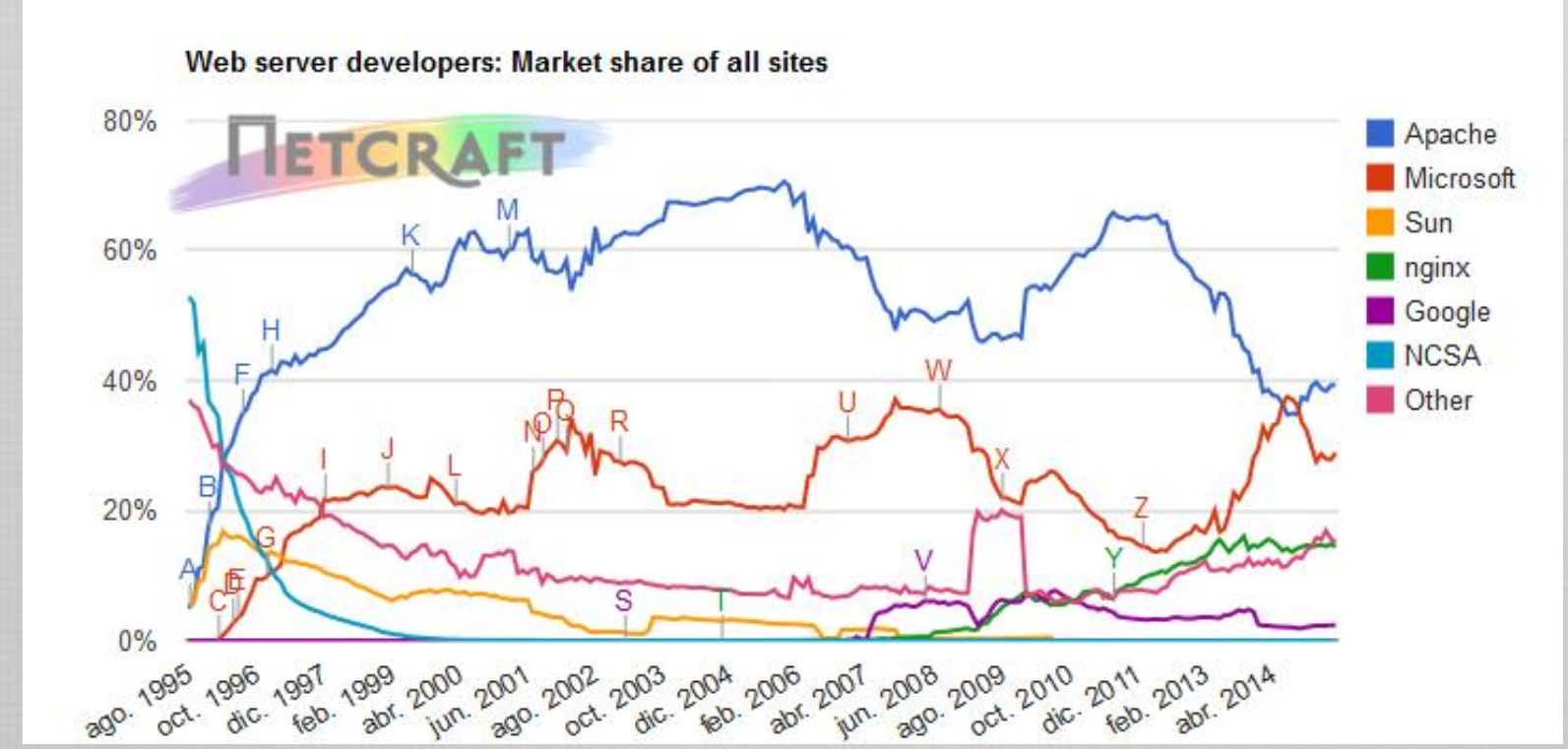
Departamento de Ciencias e Ingeniería de la Computación
Universidad Nacional del Sur

El tamaño de la web

Aproximadamente 50 billones de páginas en la web



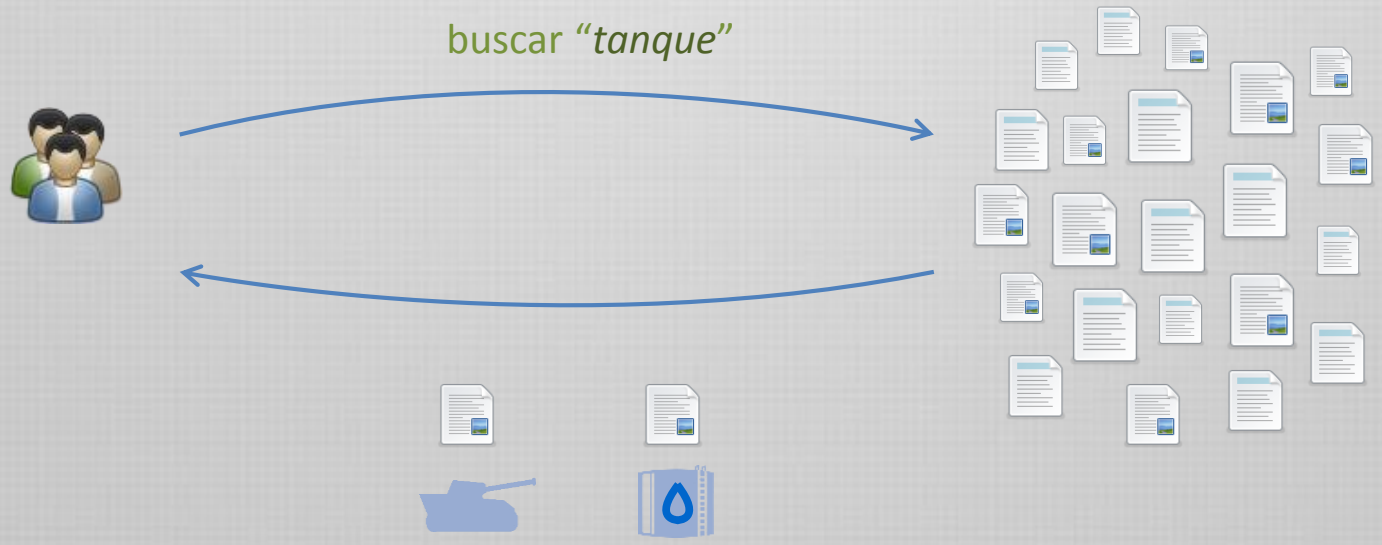
El tamaño de la web



El crecimiento de la Web

El constante crecimiento de la información en la red requiere de métodos y tecnologías especiales para individualizar las piezas relevantes.

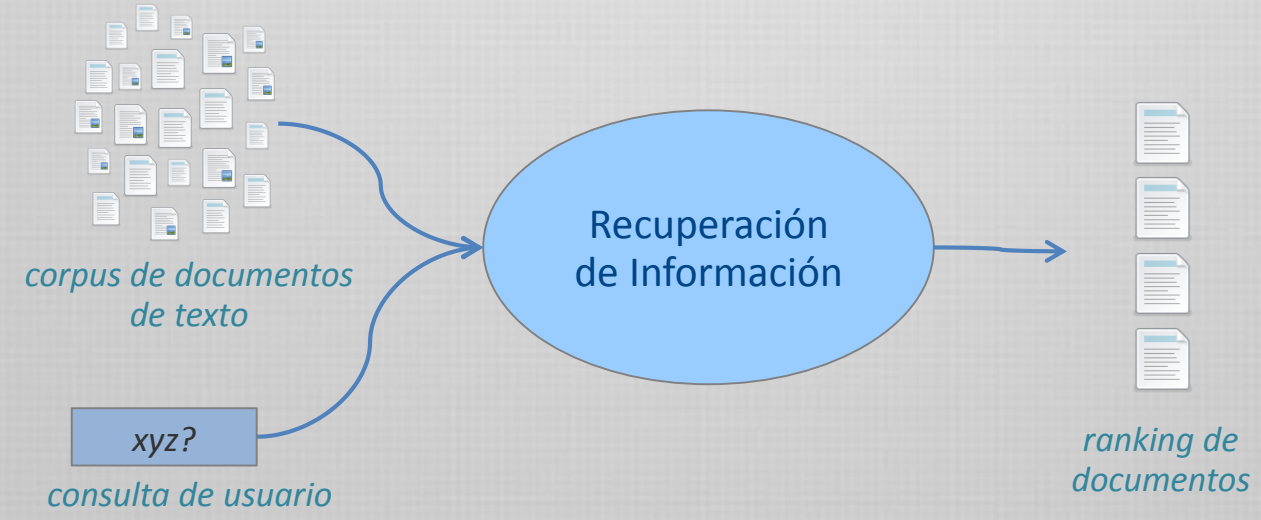
El principal inconveniente es la falta de estructuración de la información.
Es fácil de obtener, pero no fácil de interpretar.



Recuperación de Información

Recuperación de Información (Information Retrieval) es un área importante en Ciencias de la Computación, desde la década del 60.

Information Retrieval (IR) es la búsqueda y extracción de material (usualmente documentos) de grandes colecciones de datos de naturaleza no estructurada (usualmente texto) que satisface una necesidad de información.



Cobra especial interés ante la abrumadora cantidad de información en la web

Recuperación de Información - ejemplo

Brutus
AND
Caesar
AND
NOT Calpurnia

Obras Completas de Shakespeare

Una posible solución: *grepping*

- *Comenzar desde el principio, recorriendo todas las obras, marcando aquellas que poseen el término Brutus y Caesar, excluyendo los que poseen Calpurnia*

Otra posible solución: *indexing*

- *Indexar los documentos con anticipación.*
  Por ejemplo, utilizando matrices de adyacencia...

	Hamlet	*Julio César*	*Othello*
Brutus	*1*	*1*	*0*
Caesar	*1*	*1*	*1*
Calpurnia	*0*	*1*	*0*
Cleopatra	*0*	*0*	*0*

Brutus AND Caesar AND NOT Calpurnia → *1100101.. AND 111000... AND 101011..*

Recuperación de Información

Las búsquedas no siempre son tan simples

- El procesamiento debe ser rápido.
  En algunos casos, como la Web, es impracticable recorrer exhaustivamente la colección completa de datos (corpus)
- Deben permitirse operaciones más flexibles.
  Por ejemplo, buscar "celos" CERCA de "Othello", donde CERCA puede definirse previamente (x palabras, párrafos, etc)
- Deben permitirse respuestas ponderadas.
  En algunos casos puede ser deseable encontrar el "mejor" documento.
  O los mejores resultados en orden de importancia.
  ¿que significa el "mejor"?

Existen muchas técnicas, procedimientos y estrategias para organizar la información, indexarla, recorrerla y ponderarla.

El principal problema es identificar la información relevante

Existe mucho desarrollo puramente centrado en la recuperación de información en la web.

Búsqueda en la Web

Inicialmente la búsqueda en la web se realizaba de dos formas

- *Motores de búsqueda de texto como AltaVista, Excite e Infoseek.*
  Búsquedas basadas en palabras clave.
- *Indices y taxonomías mantenidos por humanos como Yahoo!*
  Navegación de indices jerárquicos.

Los motores de búsqueda han sido la herramienta indispensable para navegar en la web

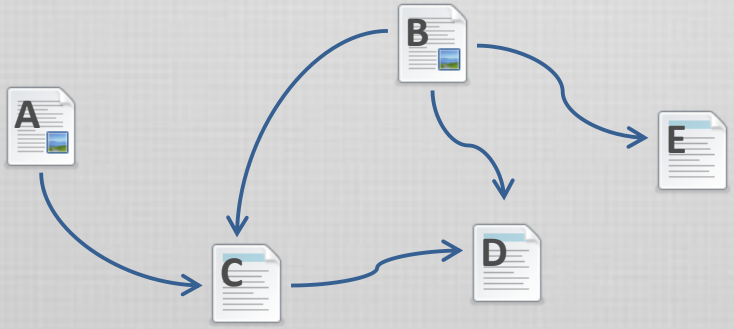
- Yahoo!
  Desde 1994, uno de los primeros buscadores. Inicialmente un directorio de páginas
- Infoseek
  Iniciado en 1994, fue un buscador importante. Los usuarios votaban los resultados.
  Luego pasó a ser Go.com. Discontinuado
- AltaVista
  Iniciado en 1995, fué un buscador muy usado. Opacado por Google.
- AlltheWeb
  Iniciado en 1999. Construía rápidamente un indice para consultas.
- AskJeeves
  Online desde 1999, todavía sobrevive. Orientado a responder preguntas de usuarios.
- Google
  Iniciado en 1998. El buscador más popular desde 2004.
- Bing
  Online desde 2009. Reemplaza a LiveSearch.

Estructura de la Web

La web es un corpus especial de documentos

- *es de gran (gran) escala,*
- *con poca coordinación en su creación y*
- *variada en contenido*

La web puede verse como un grafo donde los nodos son las páginas y los arcos son los vínculos existentes entre páginas.



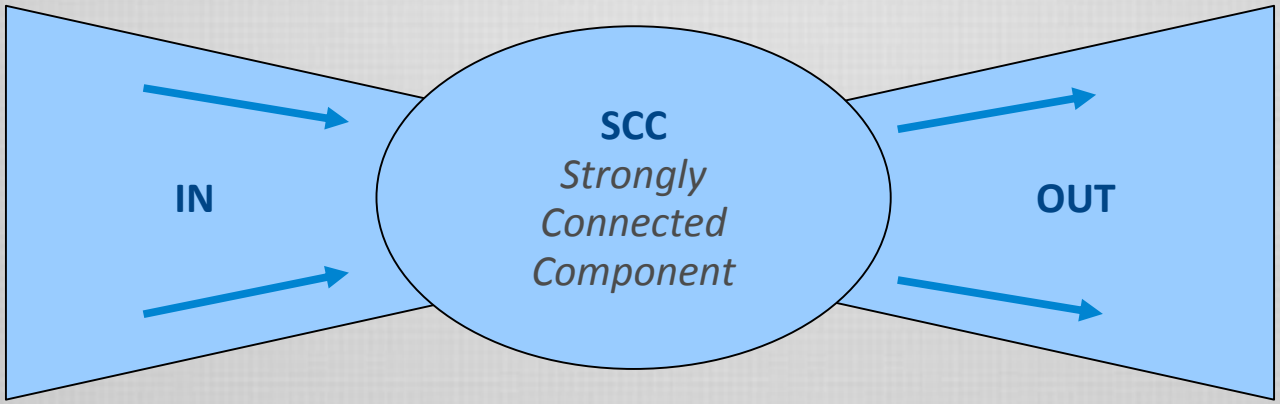
in-degree de A = 0
in-degree de D = 2
out-degree de C = 1
out-degree de B = 2

- *in-links* = hiperlinks hacia una página
  in-degree = cantidad de in-links
- *out-links* = hiperlinks desde una página
  out-degree = cantidad de out-links

Estructura de la Web

Los links no son distribuidos ni elegidos al azar.

Algunos estudios identifican ciertos patrones estadísticos y definen la estructura de la web con forma de moño (bowtie).



Páginas con links directos a páginas del SCC

Eventualmente pueden formar parte del SCC

Desde este conjunto es posible alcanzar cualquier página del conjunto

Páginas referenciadas por páginas del SCC

Búsquedas en la web

Según Andrei Broder, las búsquedas en la web pueden dividirse en tres categorías:

- Informacionales
  El usuario busca información puntual sobre un tema particular.
  No busca una página, sino varias como fuente de información.
  Por ejemplo, "Bahia Blanca", "discos portátiles", "virus H1N1"
- Navegacionales
  El usuario busca un sitio desde el cual comenzar a navegar.
  Por ejemplo, "Lufthansa".
  Usualmente ideal para "I'm feeling lucky"
- Transaccionales
  Páginas que permiten realizar actividades específicas via web.
  Por ejemplo, compras y remates on-line, bajadas de archivos o reservas de pasajes.

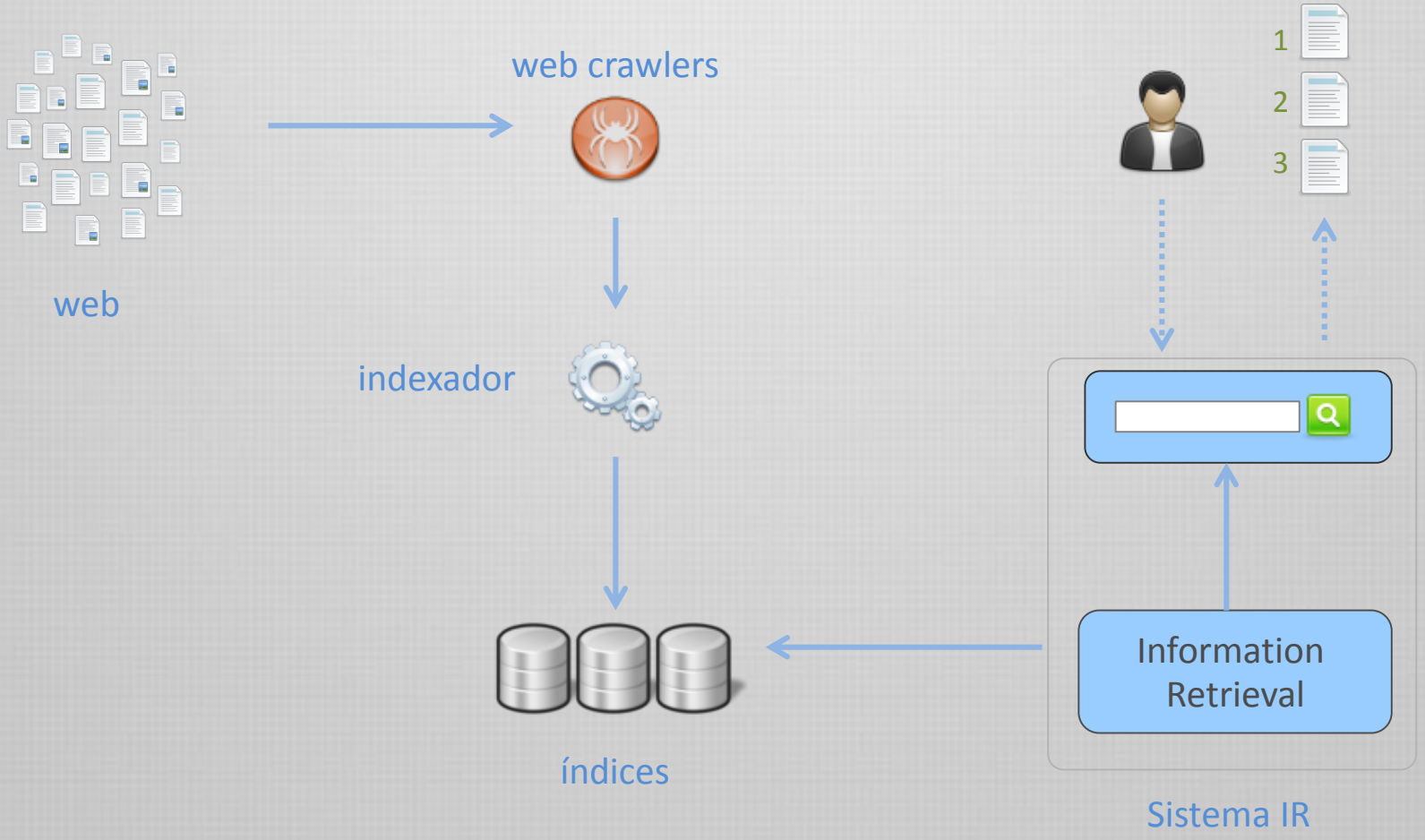
El tipo de búsqueda podría influir en el algoritmo de búsqueda utilizado.
No siempre es fácil discernir la categoría de una búsqueda.

Búsqueda web vs recuperación de información

Si bien el objetivo es el mismo (satisfacer necesidades de información), la web tiene características distintivas de la recuperación de información *clásica*.

- Tamaño
  Más de 50 billones de páginas en la web.
- Dinamismo
  La web es mucho más dinámica que los repositorios de documentos tradicionales.
- Duplicación
  Existe mucha información duplicada y redundante.
- Calidad
  Las páginas tienen todas diferente calidad visual, informativa, de accesibilidad, de actualización de datos
- Diversidad
  La variedad de tópicos es ilimitada.
- Locación
  La web está distribuida globalmente
- Reticencia del usuario
  Tiende a hacer consultas breves y simples.
- Hipertexto
  Los links no necesariamente reflejan una estructura de información relacionada.

Buscadores web



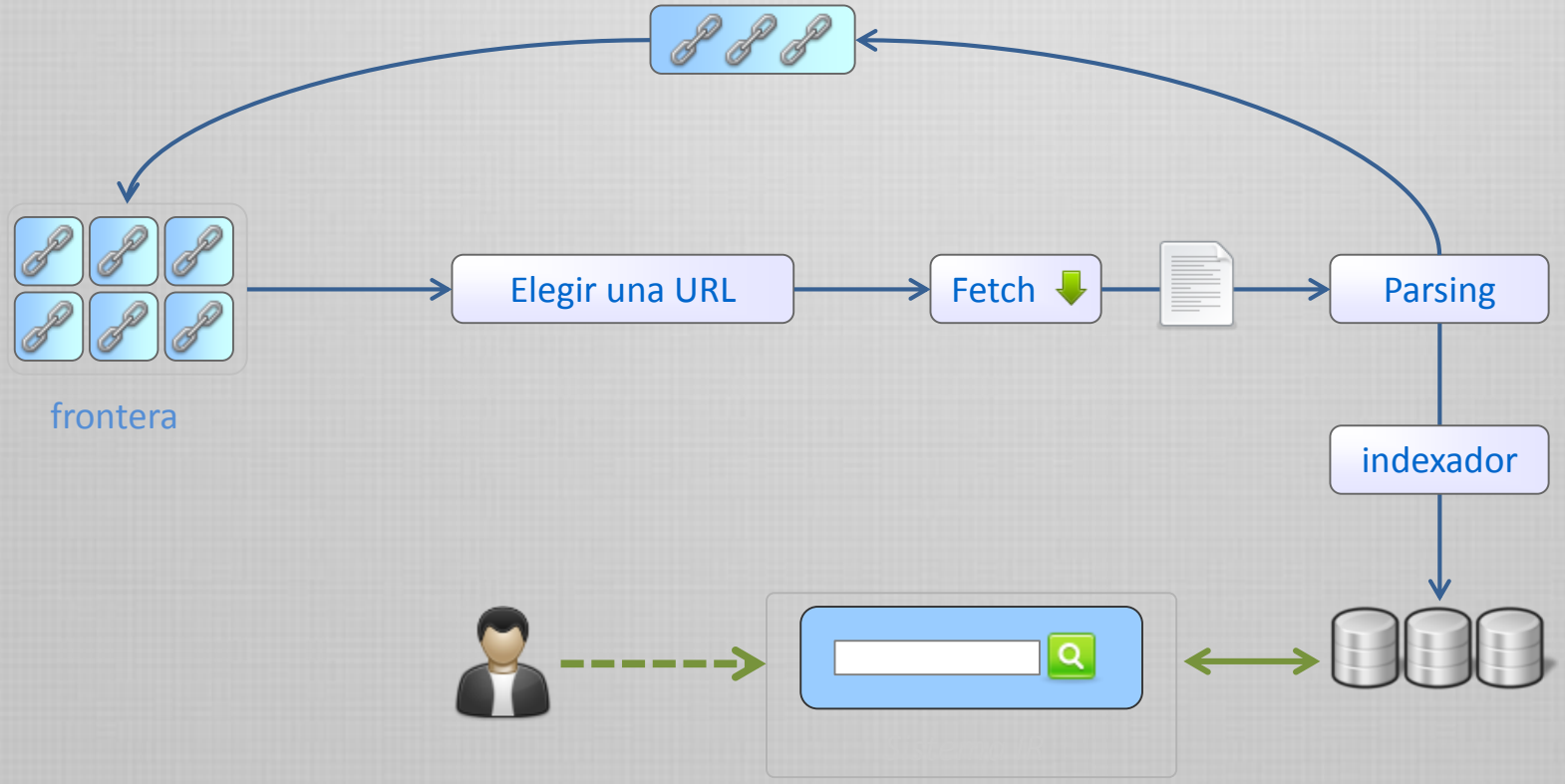
Actualmente el índice de Google
ocupa más de 100 millones de gigabytes

Web crawlers

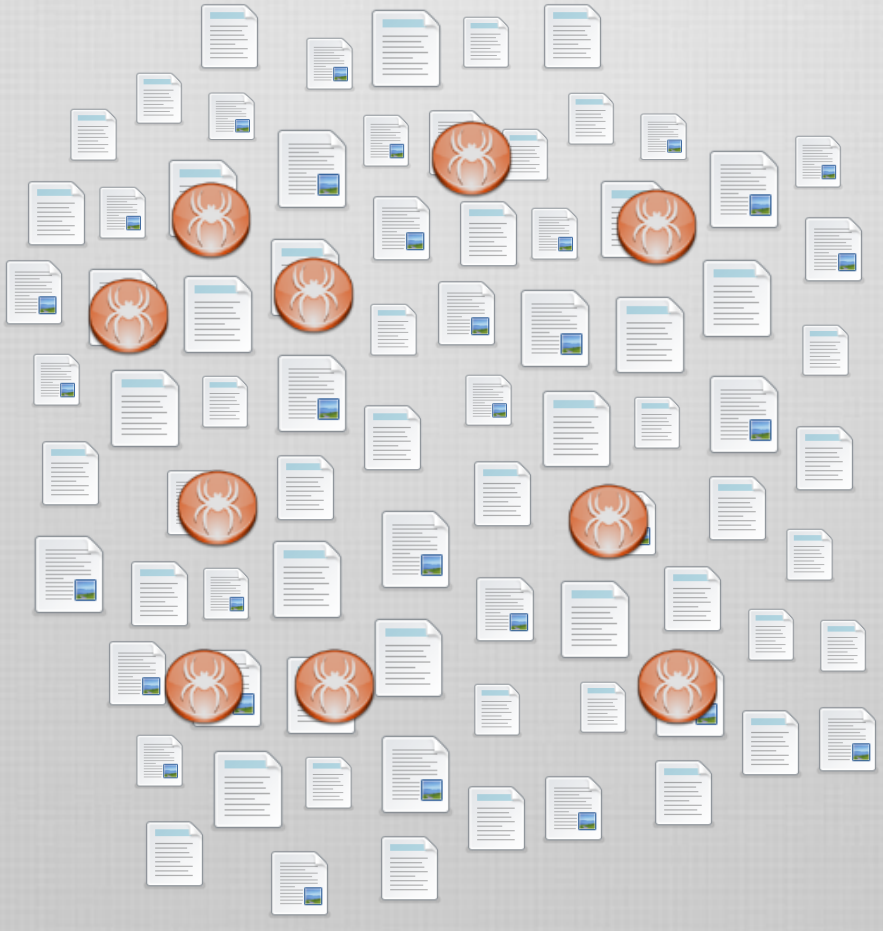
El *web crawler / spider / bot* es un programa cuyo objetivo es

- recopilar páginas web,
- para indexarlas
- sirviendo un motor de búsqueda.

Sigue la estructura de la web, rastreando páginas según los links que encuentra.



Web crawlers



Crawler simple en PHP

```
function crawl_page($url, $depth = 5) {
        static $seen = array();
        if (isset($seen[$url]) || $depth === 0) { return; }
        $seen[$url] = true;
        $dom = new DOMDocument('1.0');
        $dom->loadHTMLFile($url);
        $anchors = $dom->getElementsByTagName('a');
        foreach ($anchors as $element) {
           $href = $element->getAttribute('href');
           if (0 !== strpos($href, 'http')) {
                $path = '/' . ltrim($href, '/');
                if (extension_loaded('http')) {
                   $href = http_build_url($url, array('path' => $path)); }
                else {
                   $parts = parse_url($url);
                   $href = $parts['scheme'] . '://';
                   if (isset($parts['user']) && isset($parts['pass'])) {
                      $href .= $parts['user'] . ':' . $parts['pass'] . '@';
                   }
                   $href .= $parts['host'];
                   if (isset($parts['port'])) {
                      $href .= ':' . $parts['port'];
                   }
                   $href .= $path;
           }
        }
        crawl_page($href, $depth - 1);
}
```

Exclusión

El estándar de exclusión de robots es una convención de notificación para crawlers acerca del contenido público de ciertas páginas.

Es un protocolo netamente informativo.
El aviso se ubica en un archivo **robots.txt**

```
# robots.txt for http://www.example.com/
User-agent: *
Disallow: /blah/yadda/
Disallow: /tmp/
Disallow: /foo.html
```

```
User-agent: Googlebot
Disallow:
User-agent: googlebot-image
Disallow: /
User-agent: googlebot-mobile
Disallow: /
User-agent: MSNBot
Disallow: /
User-agent: *
Disallow:
Disallow: /cgi-bin/
Disallow: /pathprivado
Sitemap: http://misitio.com
```

Web crawlers

Propiedades que debe cumplir un crawler:

- *Robustez: evitar spider traps – páginas que confunden al crawler obteniendo infinitas páginas de un mismo lugar.*
- *Cortesía: los sitios poseen ciertas reglas sobre la visita de un spider, y esas reglas deberían observarse.*

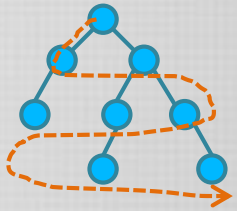
Algunos aspectos importantes:

- *Ciclos:* deben recordarse links visitados.
  La página puede referenciarse directa o indirectamente a sí misma.
- *Orden de visita*: qué página visitar primero?
  First-In-First-Out, Last-in-First-Out, Heurísticas especiales
- *Concurrencia: se puede "paralelizar" la obtención de páginas.*
  Las condiciones de la red no son constantes y pueden variar de link a link.
- *Duplicados:* la misma página con diferente URL.
  Puede asignarse a una URL un string (hash) representando el contenido
- *Calidad:* páginas malformadas con links incorrectos
  Si el crawler no encuentra lo que busca, la página se descarta.
- *Páginas dinámicas:* el contenido depende de la ejecución de (por ejemplo) JavaScript
  Puede que sean difíciles de interpretar y el crawler deba ignorarlas.

Crawlers

Basic crawler

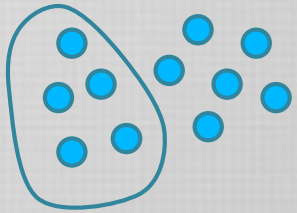
Utiliza un algoritmo básico de búsqueda (BFS, DFS)
La condición de terminación pueden ser de tiempo o espacio.



Selective crawler

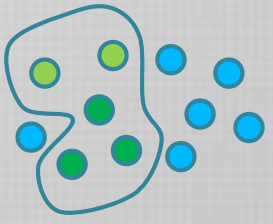
Reconoce la importancia de ciertos sitios y limita la extracción de páginas al subconjunto más importante.
La noción de importancia o relevancia puede variar

- *Profundidad, Inlinks, Pagerank*



Focused crawler

Búsqueda basada en información relacionada a ciertos tópicos, no a cualidades estructurales generales.
Realiza predicción de relevancia (e.g *clasificadores bayesianos*).
Puede incorporar aprendizaje automatizado (e.g *reinforced learning*)
Ejemplo: Citeseer



Distributed crawler

- Busca minimizar solapamiento, coordinando acciones.
- Particiona la web estática o dinámicamente.
- Requiere estrategia de particionamiento