Buscadores Web

Referencias:

http://searchenginewatch.com/

http://searchengineshowdown.com/

Curso: Recuperación de Información – 2002/1

-- ChaTo

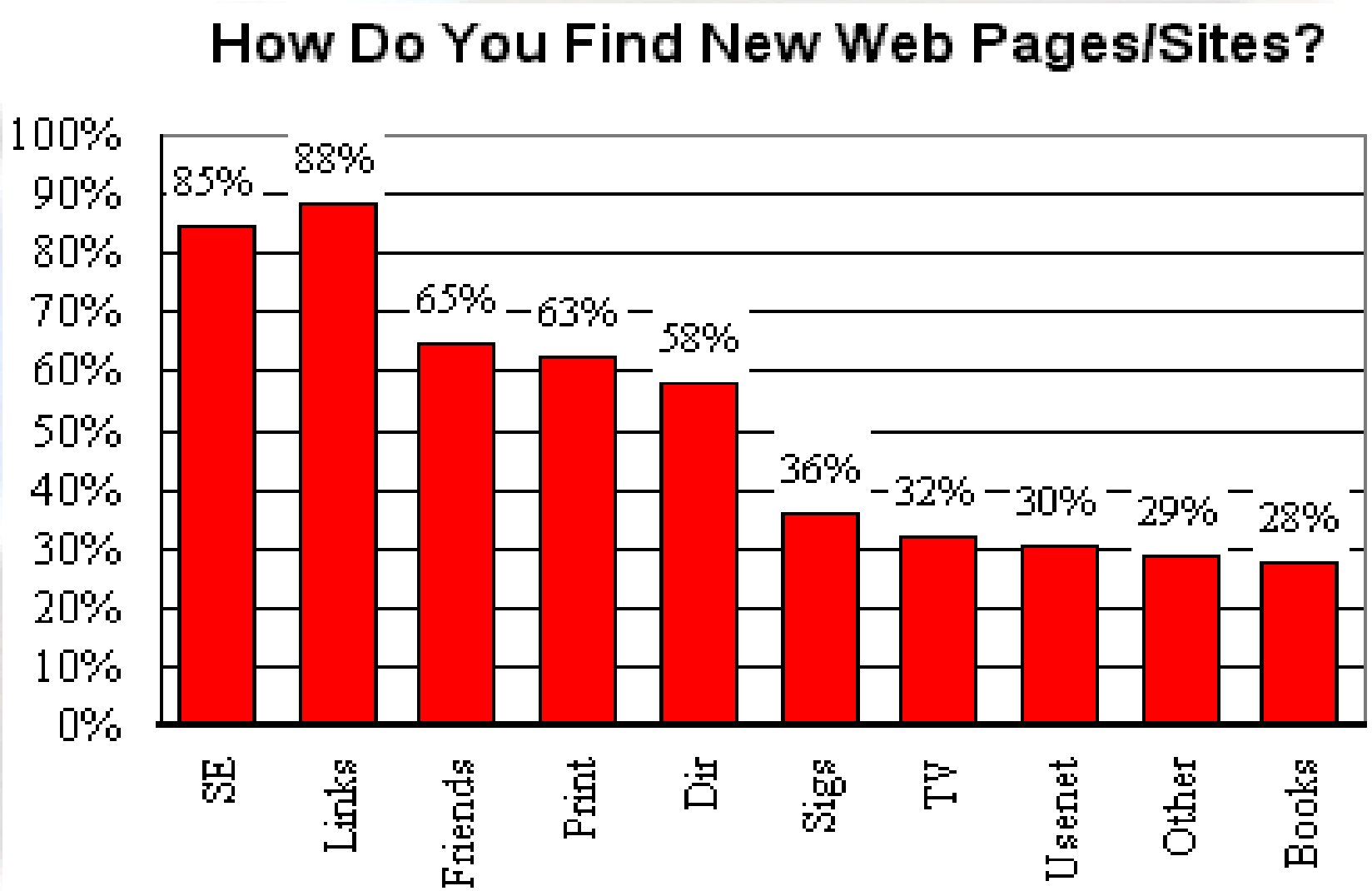
Introducción

- Sistemas para localizar documentos
  - Buscadores basados en crawlers
    - Aprox. 1M páginas por hora
  - Directorios creados por humanos
    - DMOZ recibe 250 sitios nuevos por hora

Importancia Buscadores

- Webmaster: cómo atraer tráfico al sitio
- Generan 7 a 8% del tráfico
- 97% sitios corporativos (Fortune 100) tienen problemas estructurales para máquinas de búsqueda.
- Compras en línea
  - 25% buscar
  - 5% directorio
  - 2% banner

Importancia Usuario



Importancia Usuario (2)

- 1/3 sesiones de usuario involucran buscador
- Escribir directamente la URL
  - 2001: 46%
  - 2002: 52%
- Buscar
  - 2001: 46%
  - 2002: 41%
- Buscador: descubrir nuevos sitios

Máquinas de búsqueda

- Google – google.com
  - Pagerank-based
- Yahoo – search.yahoo.com
  - Directory + Search engine
- AskJeeves – askjeeves.com
  - Natural language

- 1998, Backrub search engine.
- Múltiples buscadores especializados.
  - Catálogos, Imágenes, Noticias, Grupos, etc.
- Spellchecker, traducción, definiciones, etc.
- Adwords

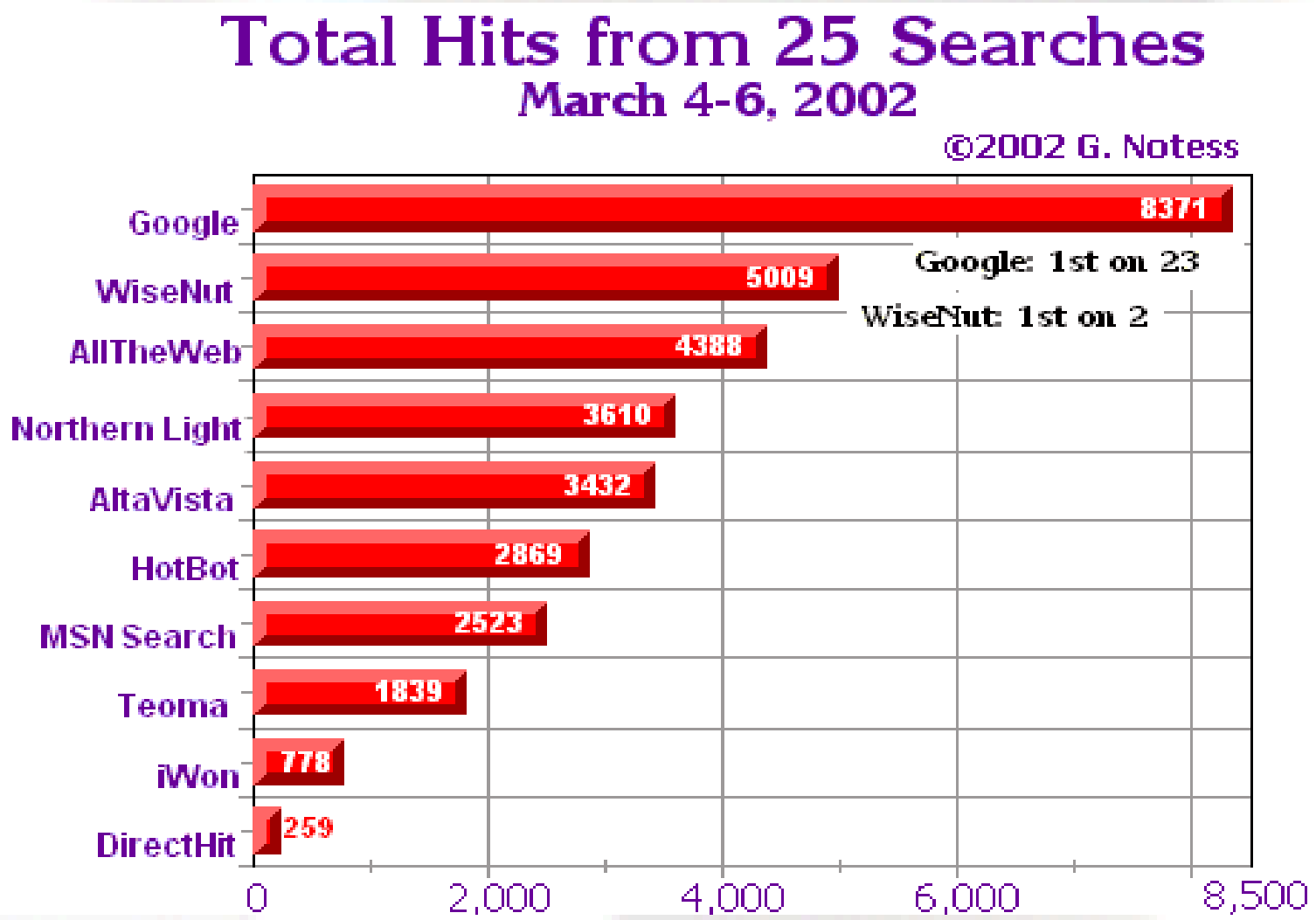
- 1994, primer directorio de páginas.
- Adquisiciones
  - GoTo, Overture, AllTheWeb, Altavista
- Máquina de búsqueda
  - Basada en Inktomi

- 1998, Primeros con lenguaje natural
  - Activo ahora en ajkids.com
- Máquina de búsqueda por Teoma

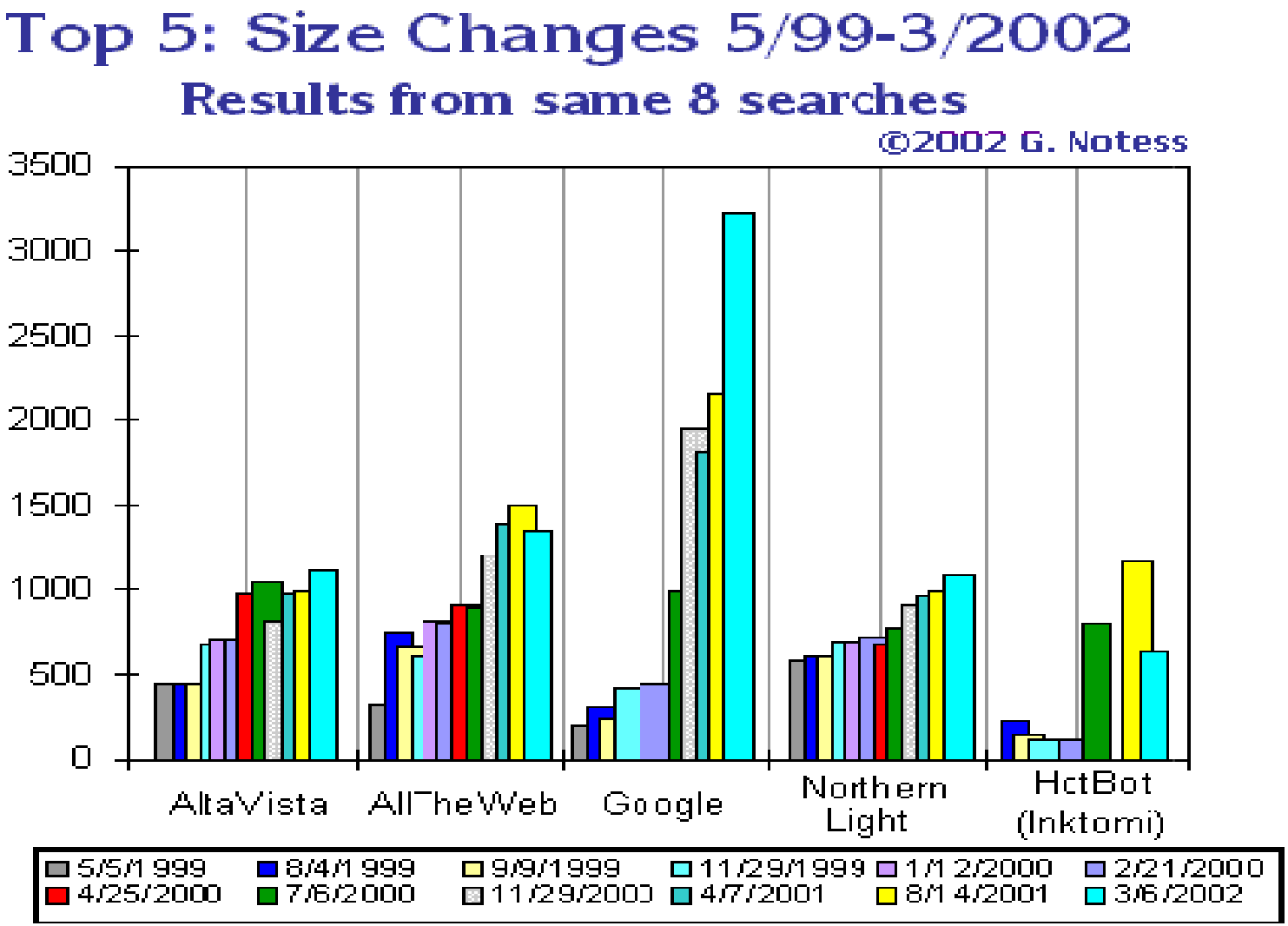
Otros buscadores

- AllTheWeb: incluye MP3 y FTP search
- Altavista: desde 1998.
- LookSmart: directorio.
- Lycos: uno de los primeros crawlers.

Comparación: Cobertura



Comparación: Crecimiento



Comparación: errores

Search Engine	% Dead	% 400 errors only
AltaVista	13.7%	9.3%
Excite	8.7%	5.7%
Northern Light	5.7%	2.0%
Google!	4.3%	3.3%
HotBot	2.3%	2.0%
Fast	2.3%	1.8%
MSN Inktomi	1.7%	1.0%
Anzwers	1.3%	0.7%

Pero (Precisión @ 5)

- Dog (Google)
  - dogpile, dog.com, hotdog, explodingdog, dogplay
- Dog (Altavista)
  - dogofday, dog.com, yellowdog, dogracing, stuffdog
- Dog (Wisenut)
  - Ilovedogs, dog.com, hotdog, dogfriendly, yellowdog
- Dog (Yahoo)
  - dog.com, dogpile, dogplay, d.o.g., explodingdog

Google (antiguo)

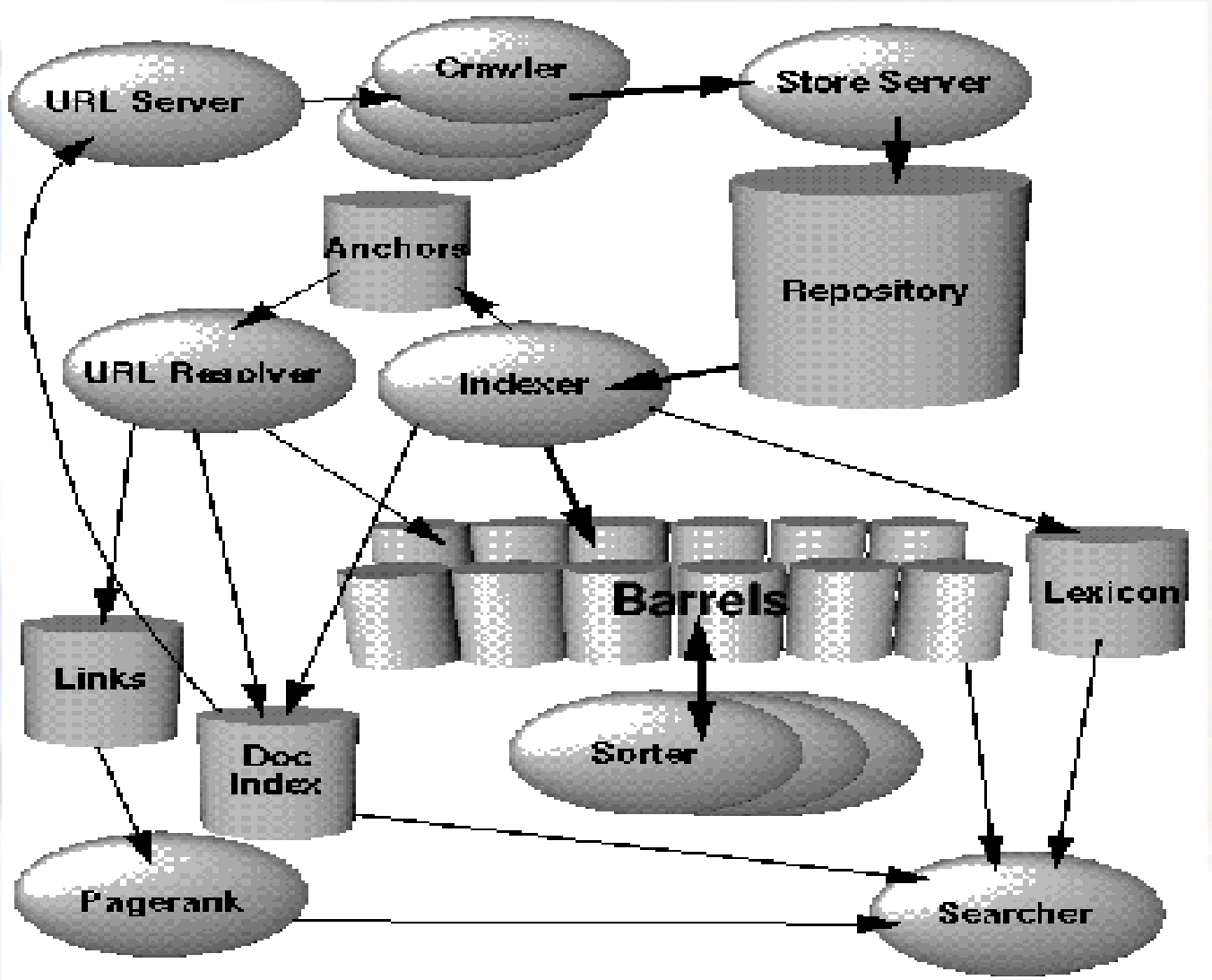
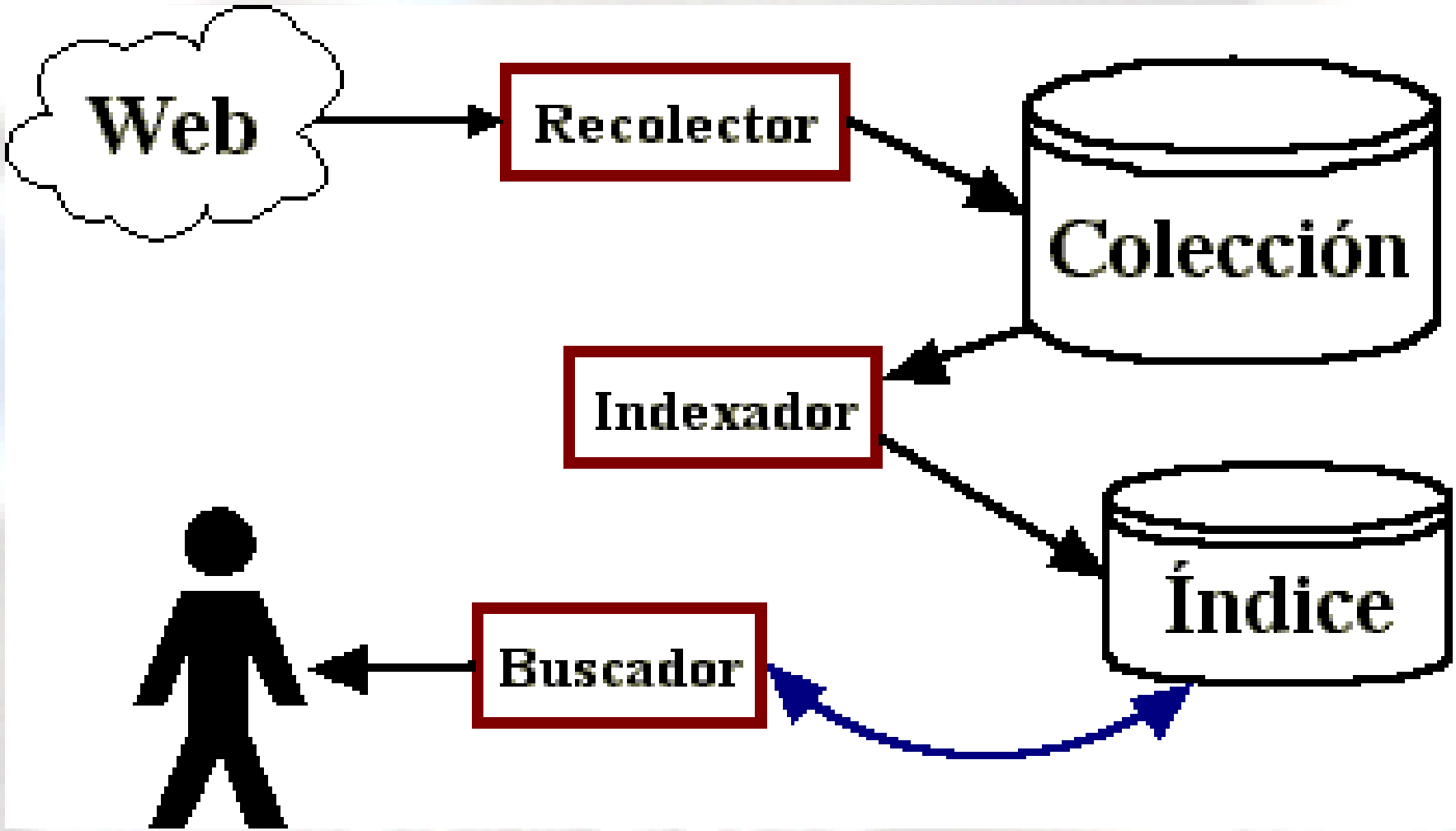
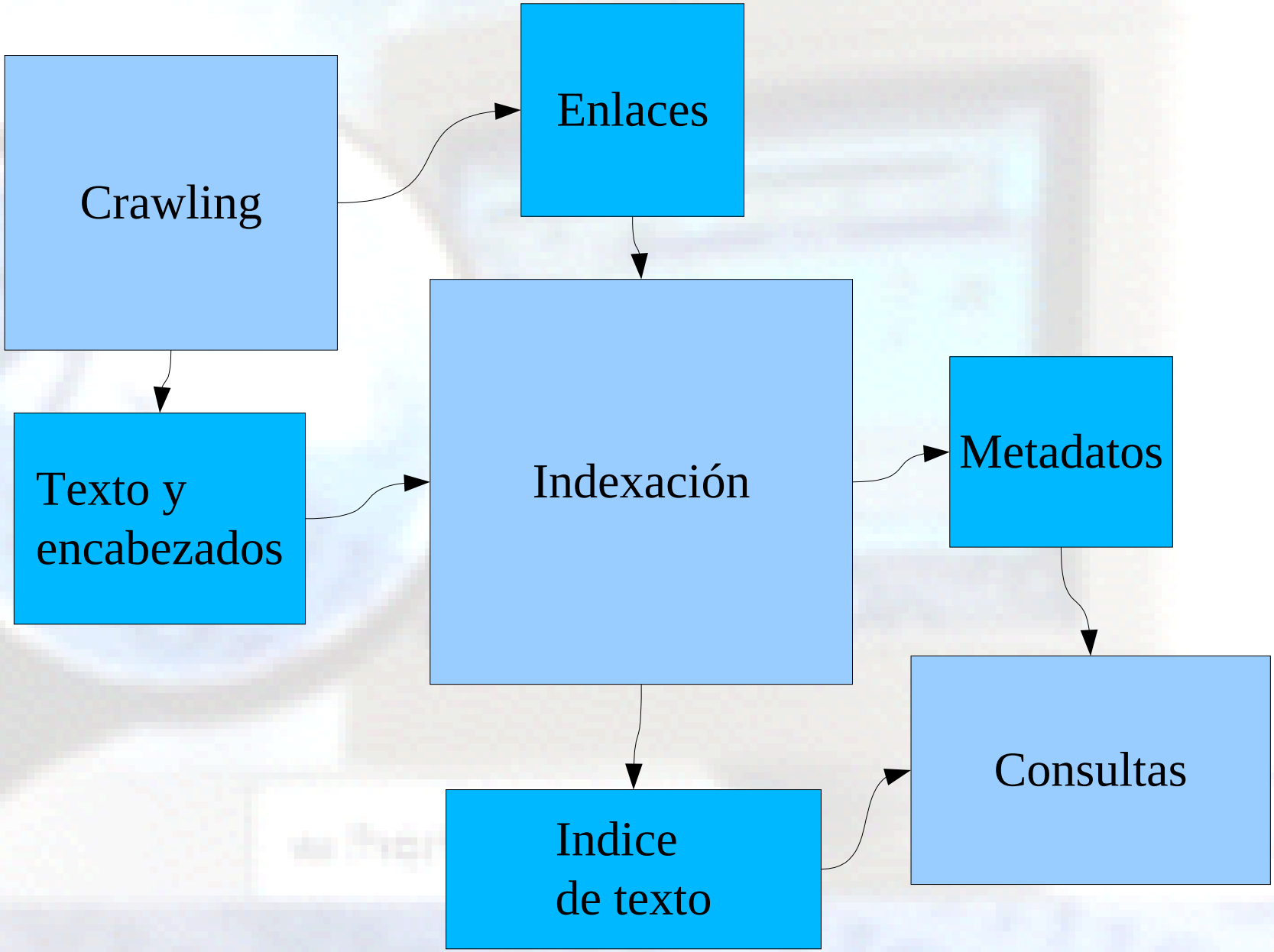


Diagrama Simple



Arquitectura general



Crawling

- Algoritmo para recorrer el grafo
- Métrica parada, Web infinita
- Parsing de los datos
  - HTML mal codificado (comillas, largo de los tags)
  - Binarios servidos como HTML
- Soporte frames
- Soporte de metatags, imagemaps
- Exclusión robots: robots.txt, meta robots
- Implementación de HTTP, etc.

Indexador

- Archivos de texto que no son HTML (PDF/PPT/etc.)
  - Toman tiempo de parsear
  - No se reducen siempre a términos (ej.: imágenes)
  - ¿Cuánto del texto indexar?
- ¿Almacenar o botar el texto completo?
  - Índice aprox. 1/3 del texto
- Indexar números
- Operaciones sobre texto: stemming
  - Eliminación de stopwords

Indexador (cont.)

- Eliminación de duplicados
  - Duplicados exactos: hashes
  - Duplicados cercanos: shingles
- Scrappers
  - Indexar definiciones
  - Indexar productos
  - Indexar imágenes/multimedios (sin bajarlos)
- Paralelización de índice
  - Por términos – Por documentos

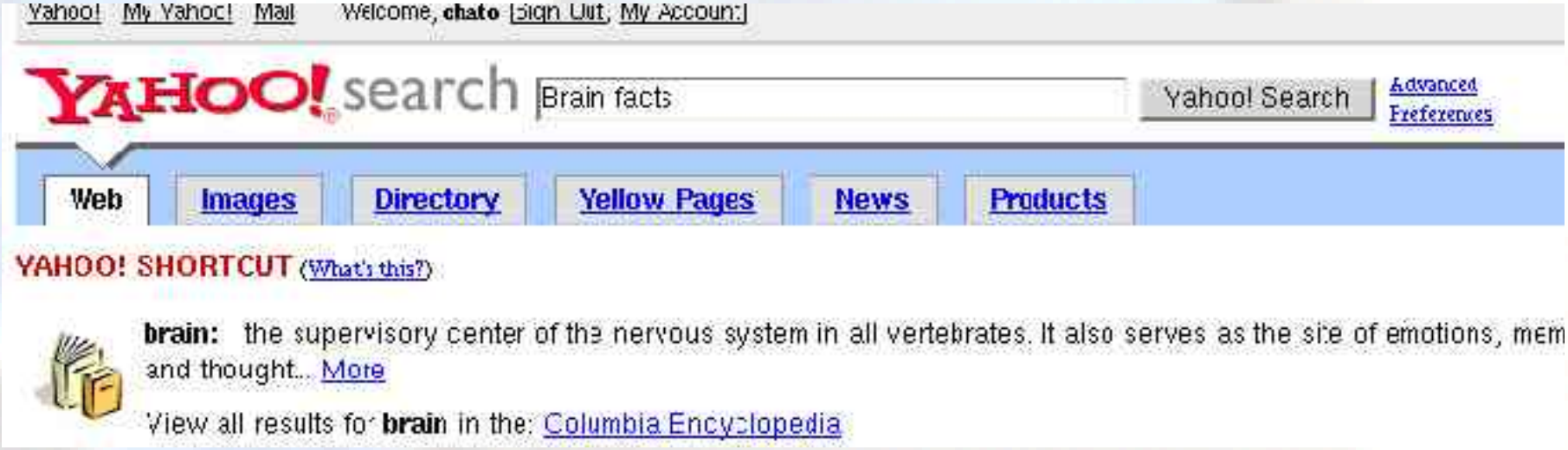
Indexador (cont.)

- Re-indexación parcial
  - Crawling con 2 o más períodos distintos
  - Combinar rankings
- Charsets
- Idiomas en que el concepto de “palabra” no está tan claro

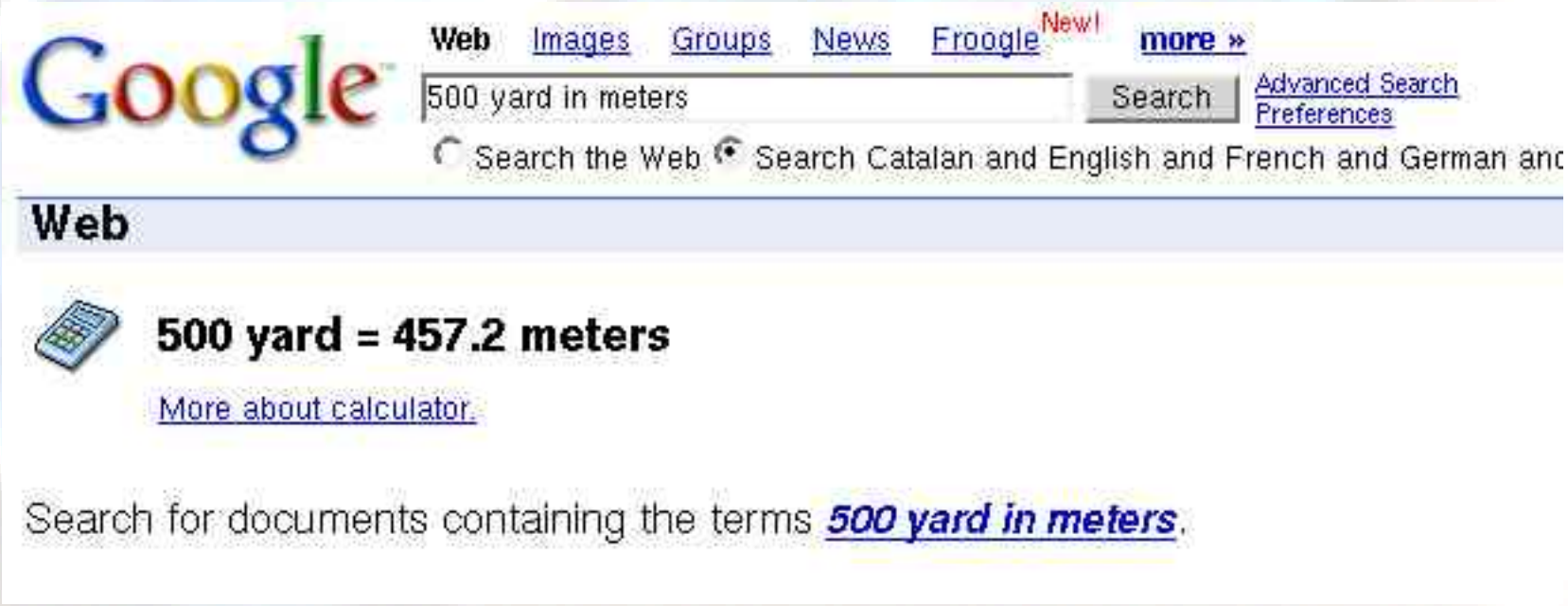
Buscador

- Operadores de búsqueda
  - Booleanos?
- Palabras “trigger”
  - Google: define, calculator
  - Yahoo: facts (encyclopedia), synonym, hotels/traffic
- Lenguaje natural

Buscador (triggers)



Buscador (especiales)



Buscador (lenguaje natural)

Buscador (corregir)

- Sugerencias de corrección
  - Muchas palabras: cuáles borrar
  - Pocas palabras: cuáles agregar
- Errores ortográficos
  - “Galiello” Yahoo: Galileo Google: Galileo
  - “Galielxo” Yahoo: - Google: Galileo
  - “Aglilloe” Yahoo: - Google: Galileo (3 errores!)
  - “Xalielo” Yahoo: - Google: Galileo
  - “Cihlxe” Yahoo: - Google: Chile

Comparación: características

Search Engines	Boolean	Default	Proximity	Truncation	Case	Fields	Limits	Stop	Sorting
Google Review	-, OR	and	Phrase	No	NO	intitle, inurl, more	Language, filetype, date	Yes, + searches	Relevance, site
Wisenut Review	-, only	and	Phrase	No	No	No	Language	Yes, + searches	Relevance, site
AllTheWeb Review	+, -, (, with)	and	Phrase	No	No	title, URL, link, more	Language, domain,	No	Relevance, site
Lycos Review	+, -	and	Phrase	No	No	title, URL, link, more	Language, domain	No	Relevance
Northern Light Review	and, or, not, (), +, -	and	Phrase	Yes * 3, auto plurals	No	title,LR_, more	Doc type, date, more	No	Relevance, site, date, folders
AltaVista Simple Review	+, -, AND, OR, AND NOT, ( )	and usually	Phrase, NEAR	Yes * 3 ** unlimited	No	title, URL, link, more	Language	No	Relevance, site
AltaVista Adv Review	and, or, and not, ()	phrase	Phrase, near, within, <, =>	Yes * 3 ** unlimited	Yes	title, URL, link, more	Language, date	No	Relevance, f Used
HotBot Review	and, or, not, (), +, -	and	Phrase	Yes *	Yes	title, more	Language, date, more	Yes	Relevance, site
MSN Search Review	AND, OR, NOT, (), +, -	and	Phrase	No	Yes	title, link	Language, date, more	Yes	Relevance
Teoma Review	+, only	and	Phrase	No	No	No	No	Yes, + searches	Folders, Relevance, metasites

Desafíos Generales

- Gran volumen de datos
  - Buscadores verticales
- Datos distribuídos
  - Problemas de red
- Datos volátiles
  - 404 Not Found
- Datos redundantes
  - 10% copias

Desafíos Generales (2)

- Datos no estructurados
  - No adhieren a estándares
- Datos de baja calidad
  - Información no confiable
- Datos heterogéneos
  - Formatos
  - Información que se desestructuro (ej.: BD->HTML)

Desafíos Específicos

- Crawler
  - DNS cuello botella
  - Velocidad variable de sitios
- Indexador
  - Conversión formatos distintos de texto
  - HTML no es respetado
- Buscador
  - Ranking: encontrar fácil, rankear difícil
  - Metabuscadores

Spamming

- URL con sessionID (oculta)
  - Diferente URL misma página
  - Variaciones ligeramente distintas
- Spamming keywords
  - Texto pequeño o transparente
  - Páginas para crawlers y páginas para humanos
- Defensa: ranking enlaces sin contar links internos
  - Ataque: varios dominios mismo dueño

Spamming (práctica)

- Texto pequeño o invisible (color similar fondo)
- Metadatos que no reflejan el contenido de la página
- Páginas que tienen redirects automáticos vía Javascript
- “Link farms”, páginas que son esencialmente links a otras páginas
- Page-swapping (página para el crawler, página para los humanos)

Spamming (práctica)

merchandise is received, please call Bob Burt at 1-321-784-6129 for a full refund or replacement.

American Flag Cases
Bob Burt
731 Java RD
Cocoa Beach, FL 32931
Phone: 1-321-784-6129
usflags@qnc.net

merchandise is received, please call Bob Burt at 1-321-784-6129 for a full refund or replacement.

American Flag Cases
Bob Burt
731 Java RD
Cocoa Beach, FL 32931
Phone: 1-321-784-6129
usflags@qnc.net

flag cases sells wooden flag cases.
veteran burials require flag cases from American flag cases.
Visit American flag cases and save on memorial flag cases!
American Honor Flag Display Cases. Memorial, Specialty, American, Honor, Flag, Display, Case, Handcrafted, Armed Forces
Custom display cases featuring memorial flag cases, saber cases, medals cases, collectibles cases, art ques, solid oak construction, wholesale and retail.

Diseño de sitios

- Títulos y meta-tags
  - Principalmente títulos adecuados
- Links encontrables
  - Etiquetas de texto cuando sea posible
  - Evitar imágenes como links
  - No ocultar links tras javascript (ej.: falsos sessionid)
- Tener buen uptime del sitio
  - DNS
  - Servidor Web

Optimización de sitios

- Sitios densamente conectados
- Todos los enlaces visibles por máquinas de búsqueda
- Todas las palabras relevantes mencionadas en cada página, o en los meta-tags
- Tener enlaces desde buenos sitios
- Search engine optimization ? Depende de cuánto prometan...