

Búsqueda en la Web

¿Qué necesidades y oportunidades específicas plantea la búsqueda en la Web?

¿Qué tipo de técnicas se utilizan para ello?

Búsqueda en la Web

- ◆ Contexto
- ◆ PageRank
- ◆ Crawling
- ◆ Otros aspectos
 - Spam
 - Detección de casi duplicados
 - Estimación del tamaño

Particularidades de la búsqueda Web

◆ Búsqueda en condiciones extremas

- Escala
- Volatilidad
- Calidad muy variable
- Colección desconocida

} Plantea muchos retos más allá de los fundamentos generales de IR

◆ Estructura adicional: hiperenlaces

- Algoritmos de ránking que explotan esta información
- PageRank y otros
- Logs masivos de búsqueda

◆ Un enorme sector de mercado

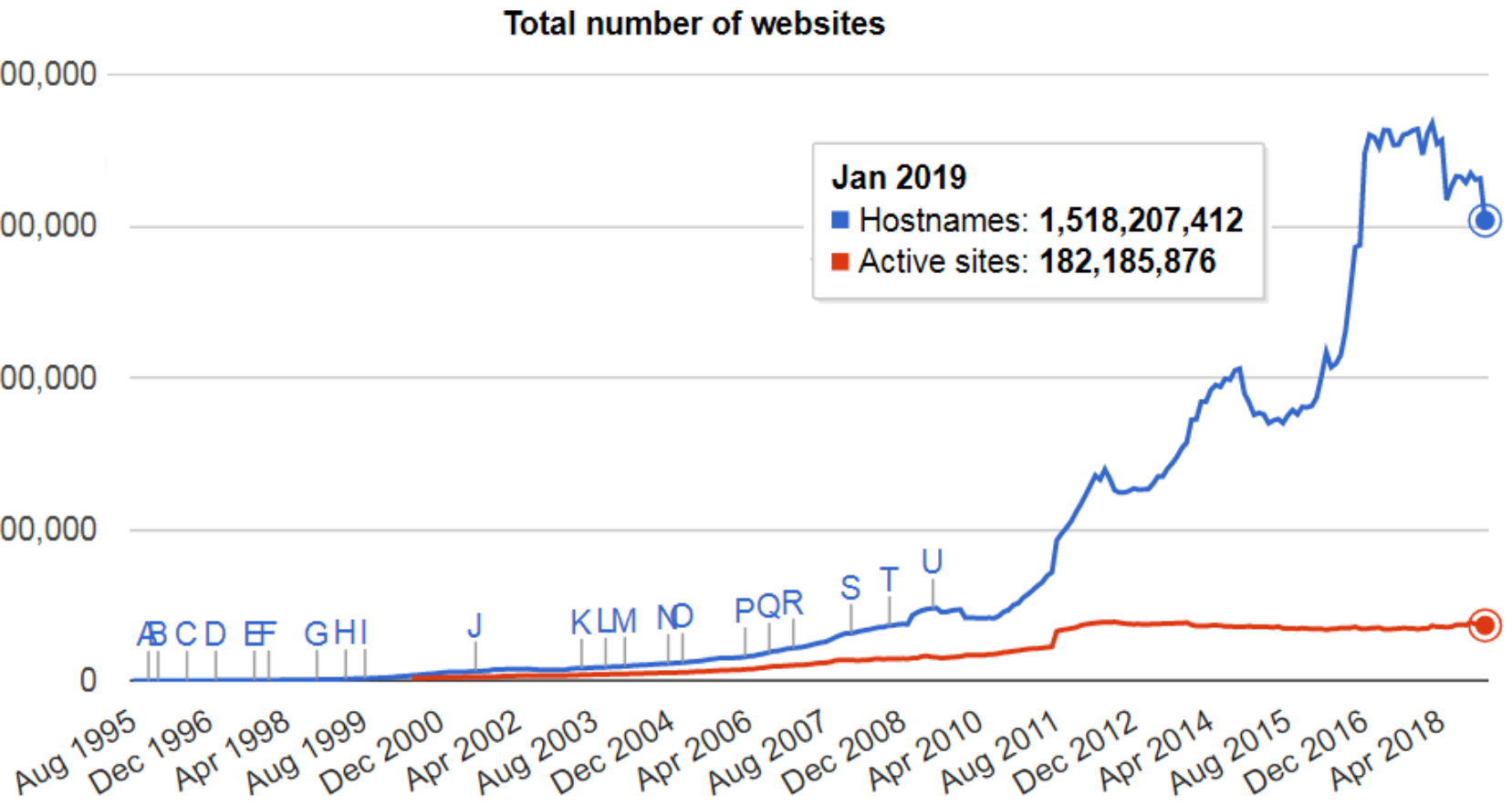
- Se rentabiliza principalmente mediante publicidad
- Ha supuesto un gran impulso al campo

Escala de la Web

- ◆ Google menciona varios billones de URLs recorridas
 - Índices del orden de cientos de PB (miles de TB)¹
 - Más la “Web profunda” no indexada
 - En realidad la Web es técnicamente infinita
- ◆ Información dispersa en cientos de millones de sitios Web
 - La indexación requiere **crawling** (semanas/meses)
- ◆ Decenas de miles de consultas por segundo (10%+ nuevas)
- ◆ Millones de resultados por consulta
- ◆ Nº usuarios: ~50% de la población del planeta
- ◆ Multimedia, multiformato, multilingüe, multidominio
- ◆ Infraestructuras masivas distribuidas

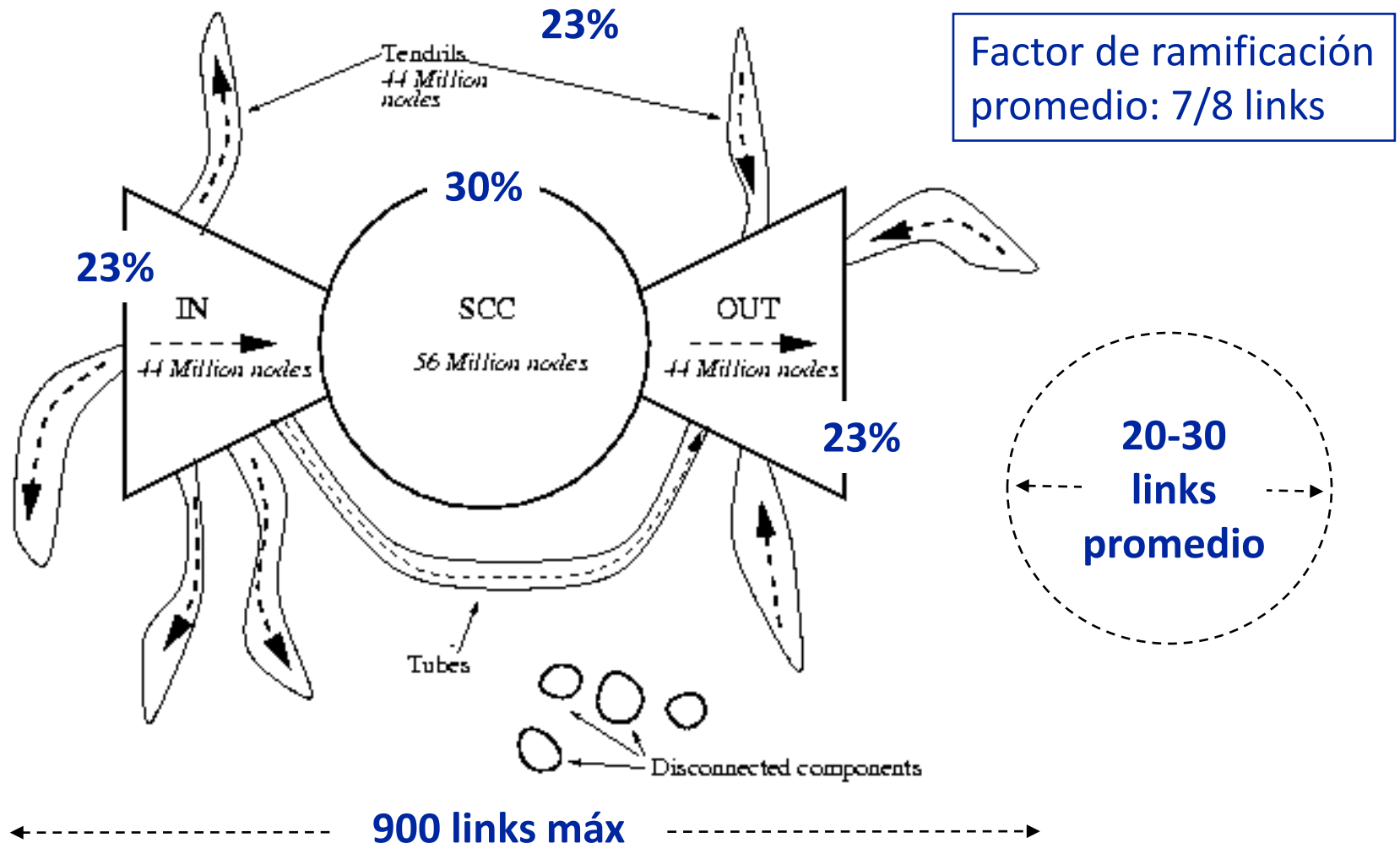
¹ <http://www.google.com/insidesearch/howsearchworks>

Escala de la Web (cont)



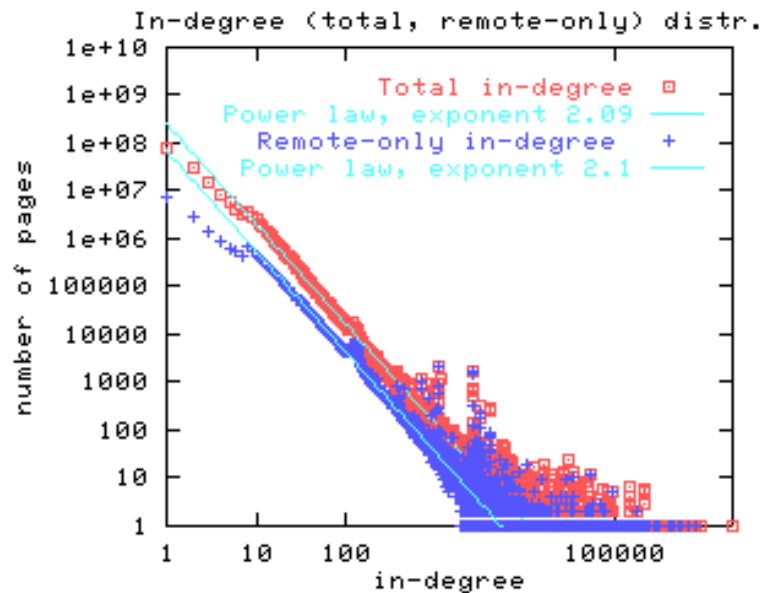
Ver <http://news.netcraft.com/archives/category/web-server-survey>

Topología macroscópica



Topología macroscópica (cont)

- ♦ Grafo libre de escala
- ♦ El nº de enlaces entrantes sigue una distribución power law
 - Unas pocas páginas concentran muchos enlaces
 - Muchas páginas tienen sólo unos pocos



Volatilidad de la Web

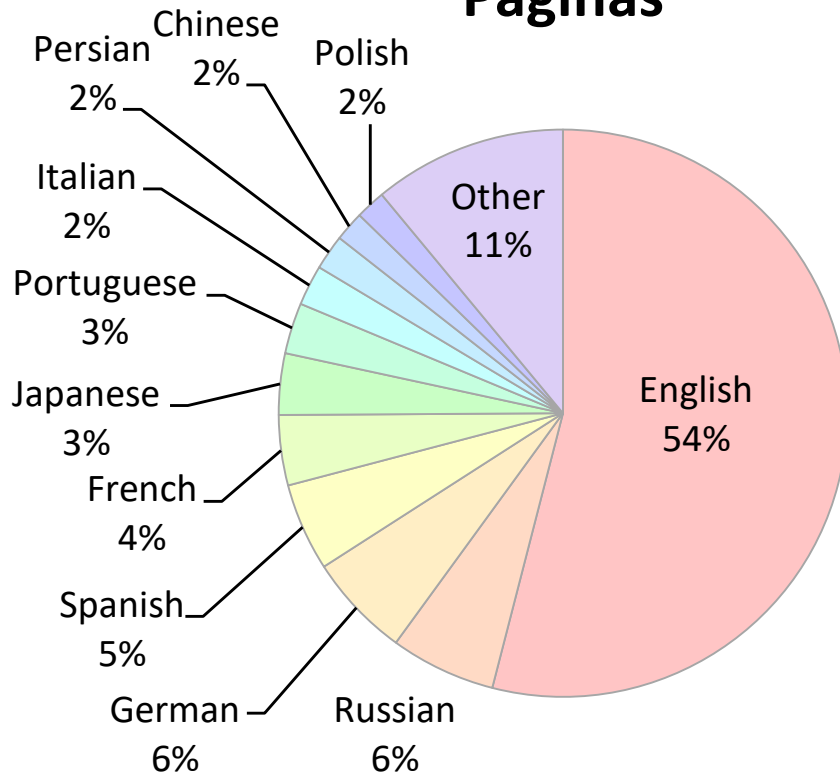
- ♦ Un alto porcentaje de la Web cambia cada mes (o cada día, o cada segundo...)
 - ~ 50% de los sitios Web desaparecen enteros de un año al siguiente según estudios
En los sitios Web aparecen y desaparecen páginas periódicamente
 - El contenido de la mayoría de las páginas se modifica con frecuencia
 - Muchas páginas son enteramente dinámicas y/o altamente cambiantes
 - Nuevos medios: blogs, mircoblogs (Twitter), foros, portales de noticias, etc., son streams dinámicos de información más que “páginas” estables
- ♦ Se precisa un **crawling continuo** para actualizar el índice constantemente

Calidad variable

- ♦ Calidad enciclopédica, calidad media, información anecdótica, texto improvisado, spam...
- ♦ Descentralización
 - Ausencia de organización global ni supervisión editorial
 - Cualquiera puede crear un sitio Web
 - Cualquiera puede aportar contenido
 - ~ 70% del contenido de la Web **generado por usuarios** finales
- ♦ **Duplicación** de contenido (mirrors, etc.)
- ♦ **Spam**
 - Autores y buscadores en posición de adversarios
 - Contenido engañoso, enlaces engañosos, promoción fraudulenta
 - Múltiples mecanismos de detección y penalización de spam

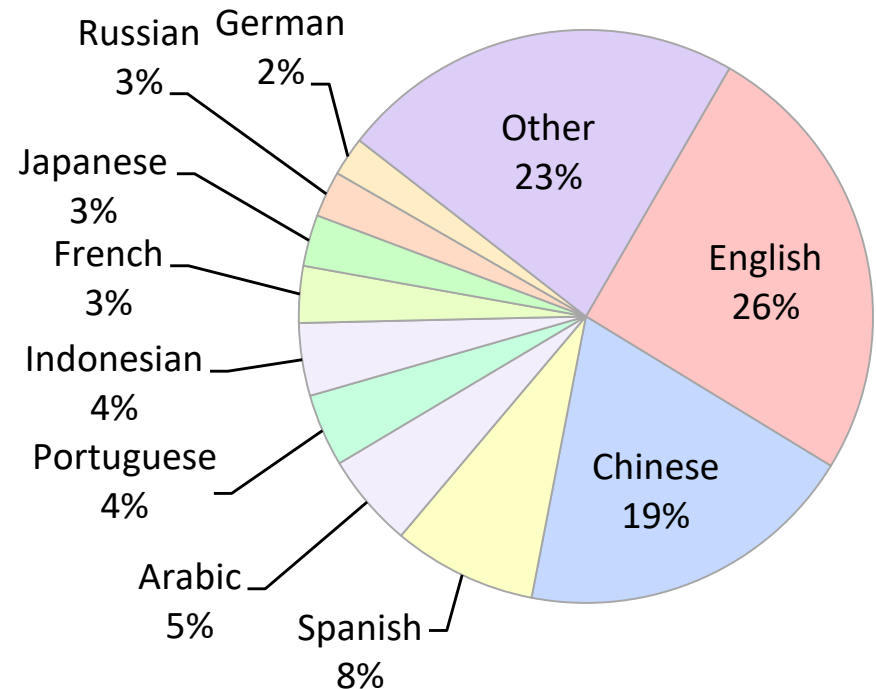
Idiomas en la Web

Páginas



http://w3techs.com/technologies/overview/content_language/all
(marzo 2019)

Usuarios



<http://www.internetworldstats.com/stats7.htm>
(diciembre 2017)

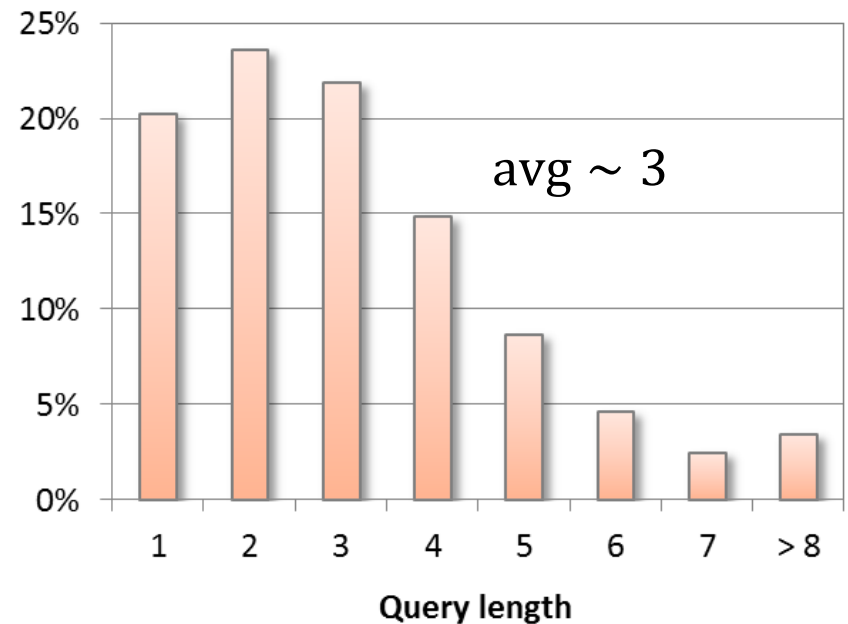
Los usuarios

◆ Perfil universal

- No se presuponen capacidades o conocimientos
- Todo tipo de dominios (búsqueda de propósito general vs. búsqueda vertical)
- Internacionalización

◆ Tendencia superficial

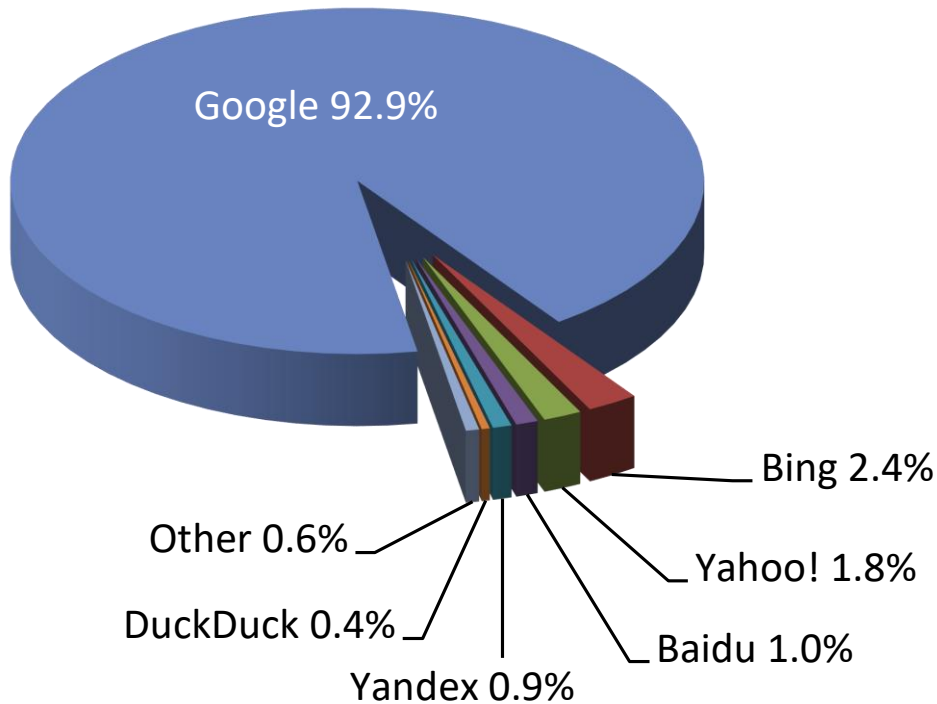
- Consultas cortas (2-3 palabras en promedio, tendencia creciente)
- Predominio de sesiones cortas
 - Browsing poco profundo
 - Pocas reformulaciones
- Búsquedas navegacionales y transaccionales frecuentes (~25-20%)



(Estudios Hitwise 2009)

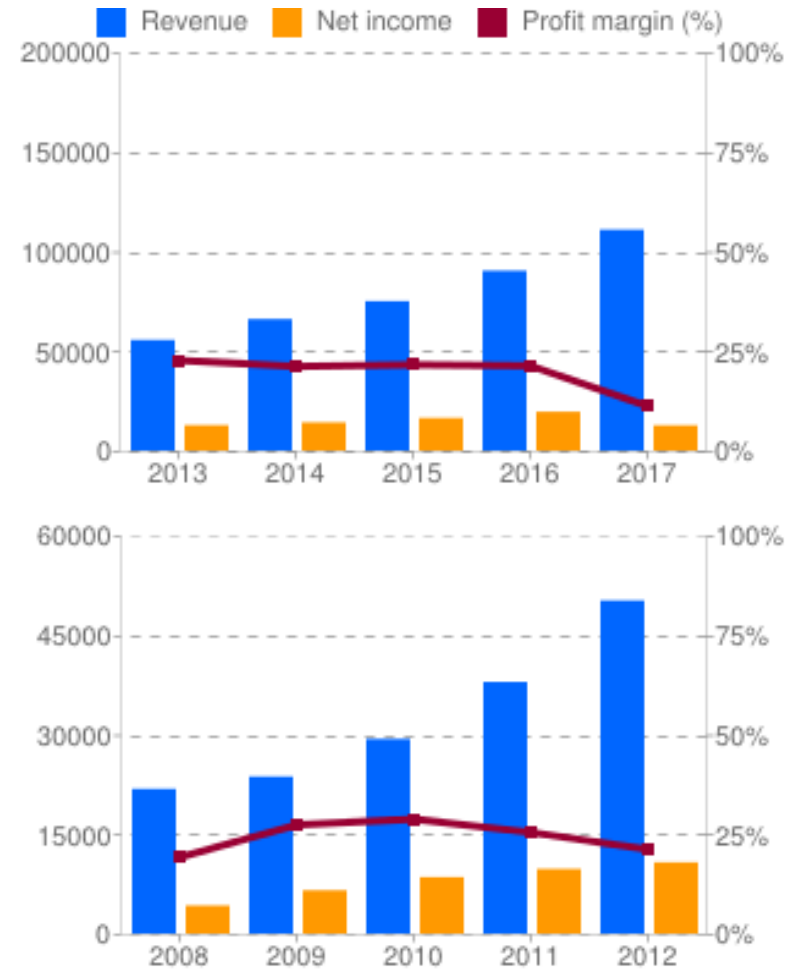
El mercado de la búsqueda Web

Tasa de mercado
(en % consultas)

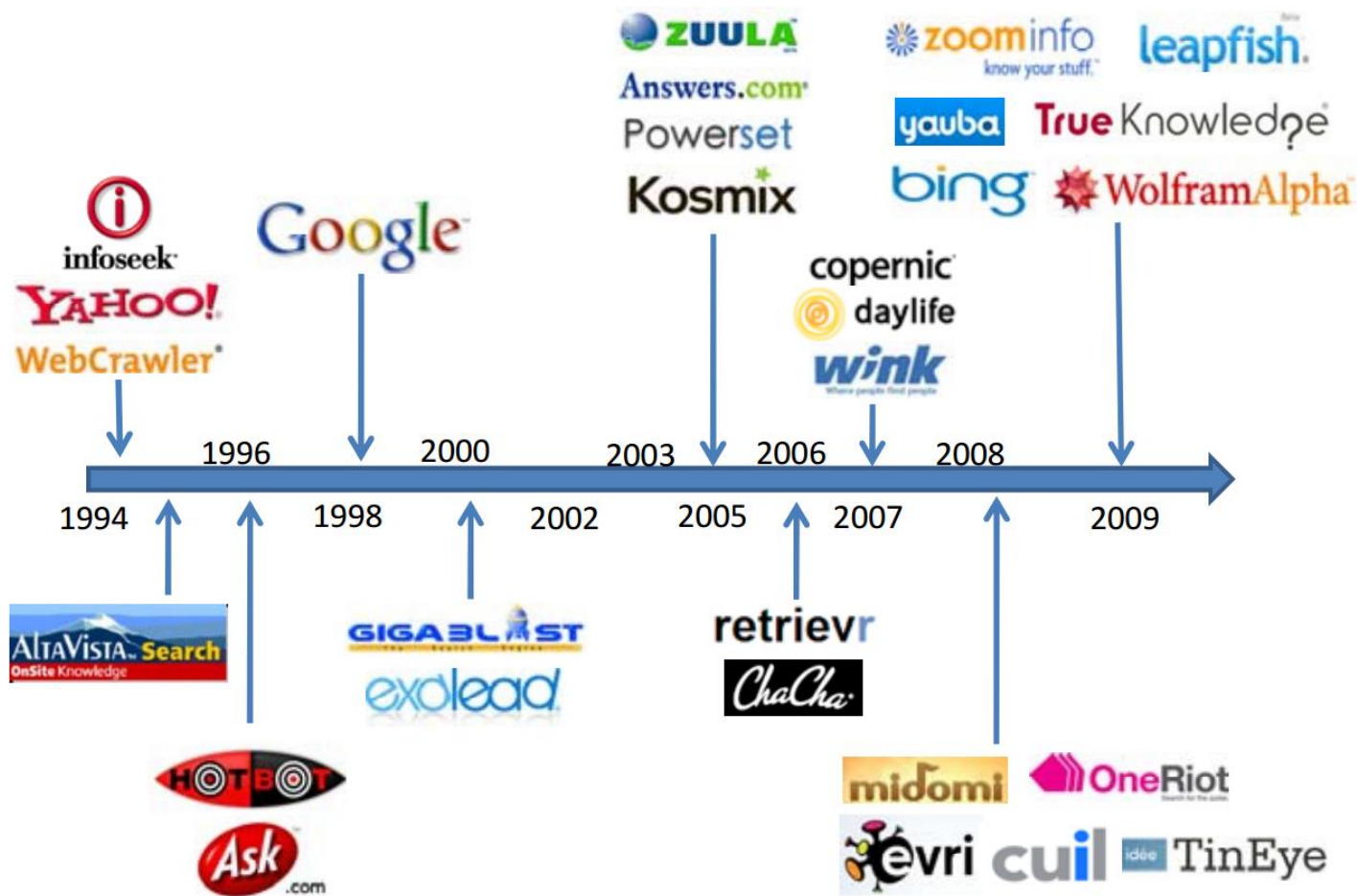


[http://gs.statcounter.com/
search-engine-market-share](http://gs.statcounter.com/search-engine-market-share)
(enero 2019)

Cuenta resultados Google



El mercado de la búsqueda Web



T. Buganza and E. Della Valle. The Search Engine Industry. In: S. Ceri and M. Brambilla (Eds.): Search Computing, LNCS 5950, 2010

Ver también: http://www.bruceclay.com/serc_histogram/histogram.htm

El mercado de la búsqueda Web

- ◆ En los 90
 - Primeros servicios basados en directorio, JumpStation (1993, títulos), WebCrawler (1994, full text), Go, Lycos, AltaVista, Magellan, Excite, Infoseek, Inktomi, Ask, Northern Light, Yandex, Google, MSN, AllTheWeb, Teoma...
- ◆ En los 00
 - Baidu, Exalead, Yahoo!, AOL, Bing...
 - Amplias adquisiciones y fusiones
- ◆ Rápida hegemonía de Google desde principios de los 00
 - Efectividad (relevancia), eficiencia, sencillez, calidad de producto, diversificación, filosofía corporativa...
 - Introduce el sistema de resultados patrocinados (AdWords, AdSense, pay per click...)
 - Hoy ~100.000 empleados, ~140.000M\$ ingresos anuales
- ◆ SEO
 - \pm 90% de los nuevos visitantes de un portal llegan desde un buscador
 - “Ingeniería inversa” del ranking de los buscadores: técnicas, herramientas, perfil profesional
- ◆ Fusión con más aplicaciones, servicios y tecnologías
 - Q&A, entity search, diccionario, calculadoras (conversiones, plots, etc.), traducción, imágenes, vídeo, mapas, libros, literatura científica, meteorología, información financiera, resultados deportivos...

El éxito de Google

- ♦ Calidad de servicio, sencillez, cobertura
- ♦ Organización horizontal
- ♦ Entorno colaborativo y abierto
- ♦ Captación y retención de talento
- ♦ Meritocracia
- ♦ Visión “filantrópica”

- ♦ Expansión            ...

“You can make money without doing evil”

“You can be serious without a suit”

“Work should be challenging and the challenge should be fun”

“We believe strongly that in the long term, we will be better served —as shareholders and in all other ways—by a company that does good things for the world even if we forgo some short term gains”

Evolución de Google



Jeff Dean
Senior Google
Fellow

- ♦ En 1999 Google indexaba 50.000 páginas en 1 mes
 - Diez años más tarde lo hacía en menos de 1 minuto
- ♦ Índice distribuido
 - Particionado por documentos
 - Replicaciones del índice para paralelización
 - Múltiples métodos de compresión: Huffman, VBC, δ , Rice...
 - Soluciones in-memory
- ♦ Una consulta se procesa en miles de máquinas
- ♦ En torno a la mitad de las consultas se responden por cache
- ♦ Ver más en <http://static.googleusercontent.com/media/research.google.com/en//people/jeff/WSDM09-keynote.pdf>

Google hoy - Infraestructura

Google Data Centers
(Billones de documentos)



Data centers

Georgia



Data centers

Oregon



Data centers

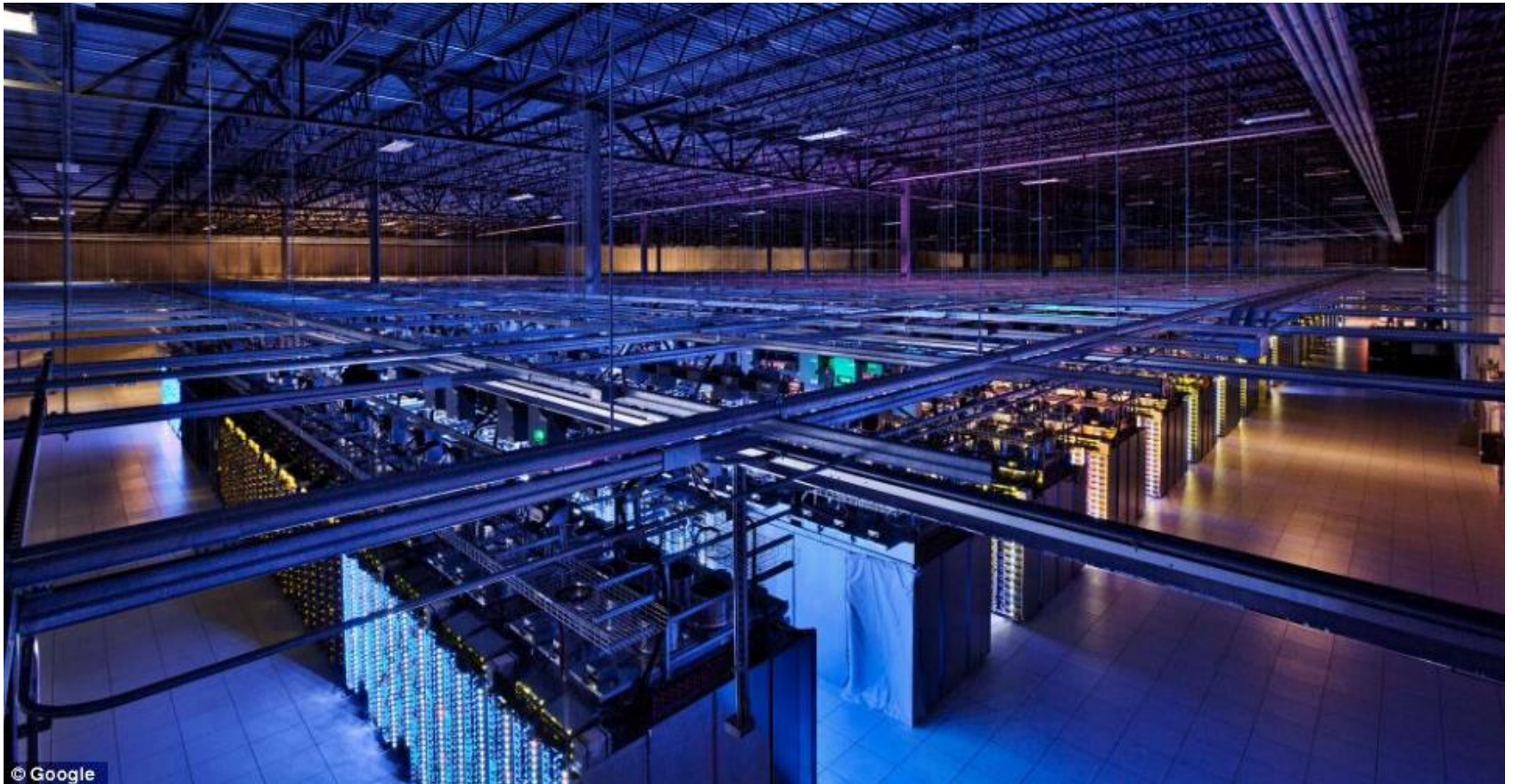
Finlandia



<http://www.google.com/about/datacenters>

Data centers

lowa (clusters)



<http://www.google.com/about/datacenters>

Data centers

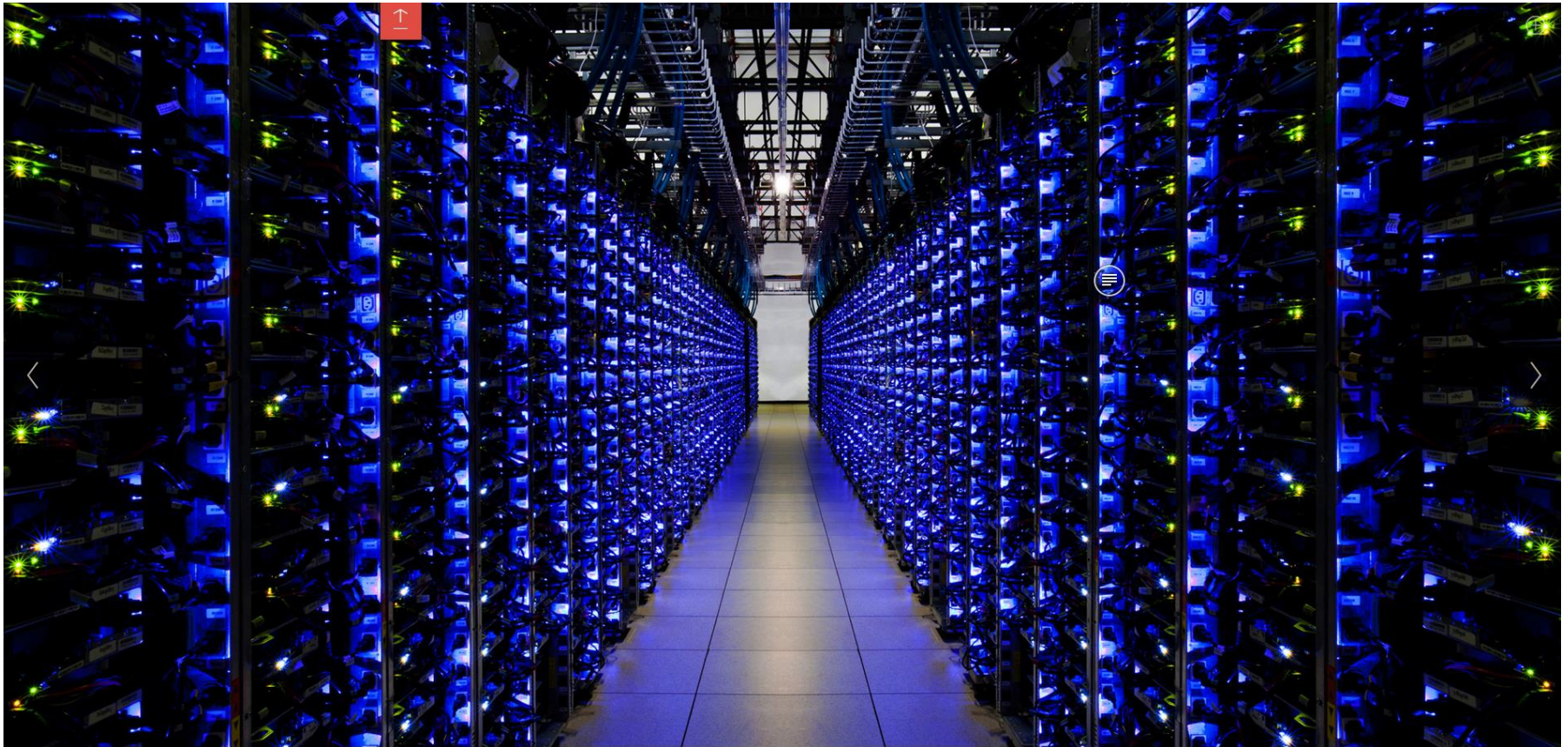
Oklahoma (clusters)



<http://www.google.com/about/datacenters>

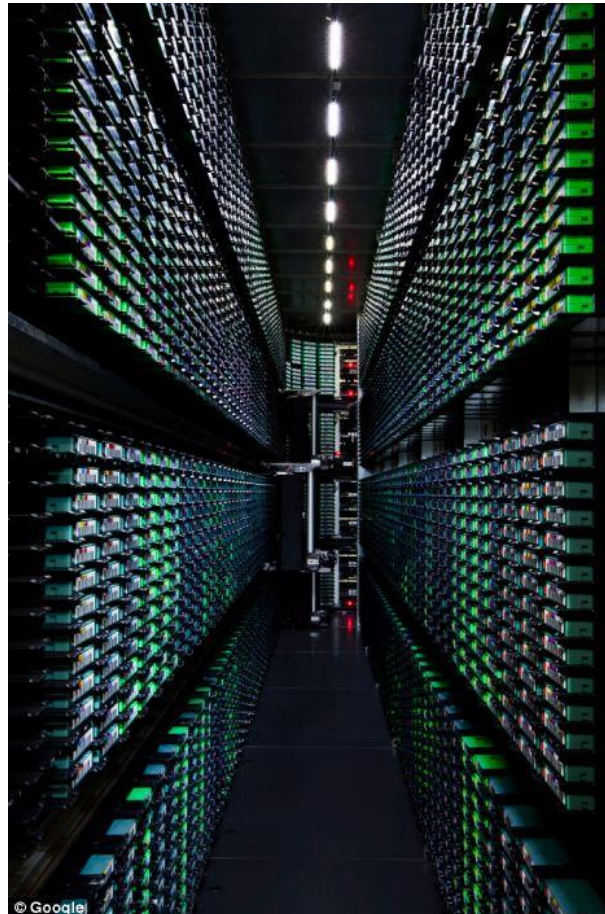
Data centers

Georgia (clusters)



Data centers

Berkeley (backup en cinta)



<http://www.google.com/about/datacenters>

Data centers

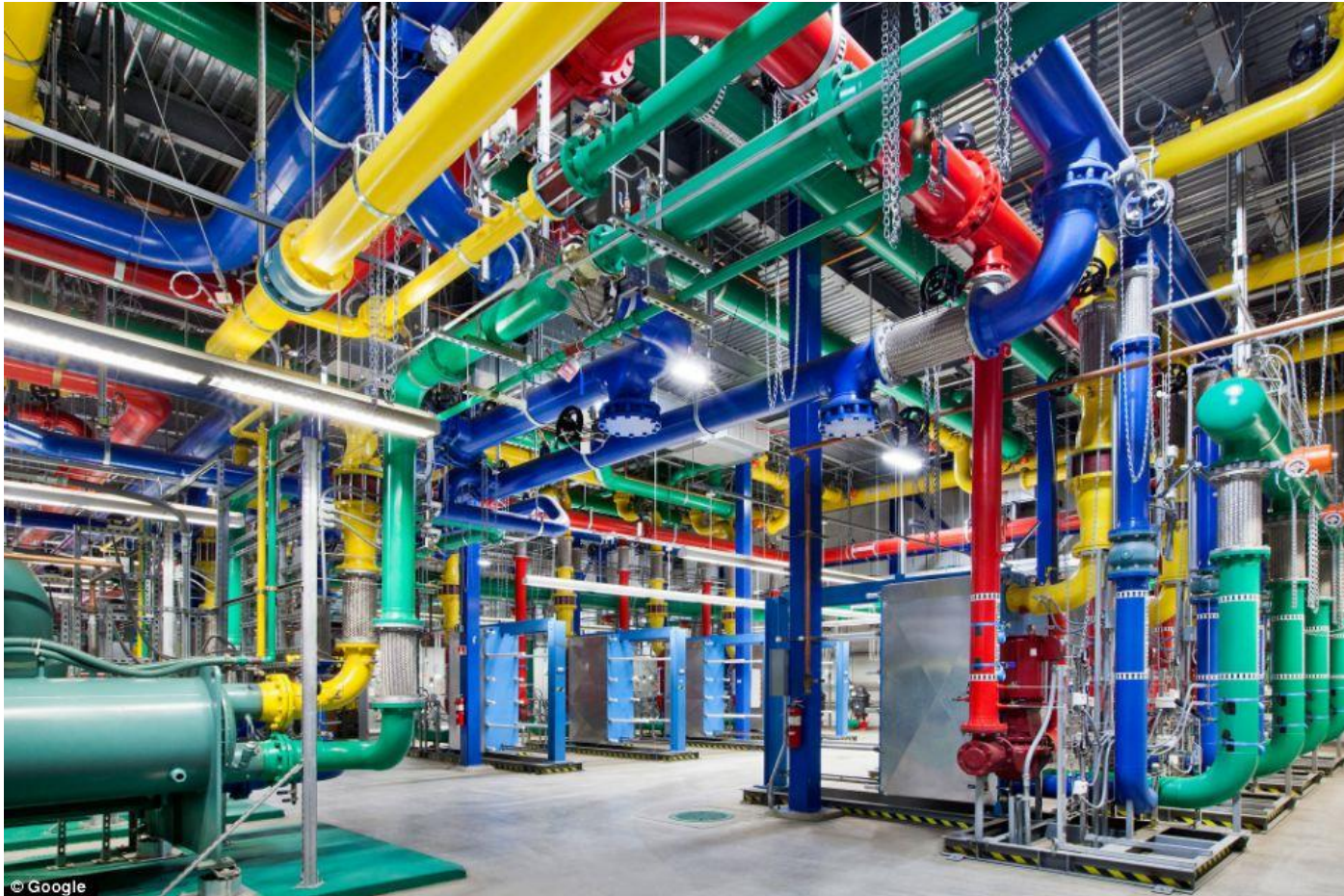
Oklahoma (refrigeración)



<http://www.google.com/about/datacenters>

Data centers

Oregon (refrigeración)



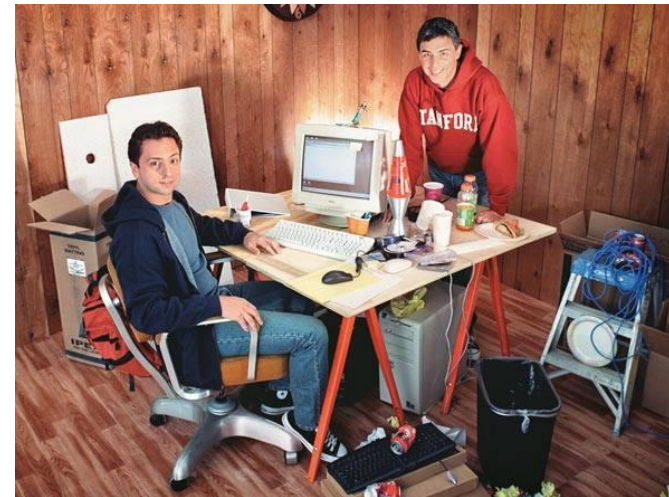
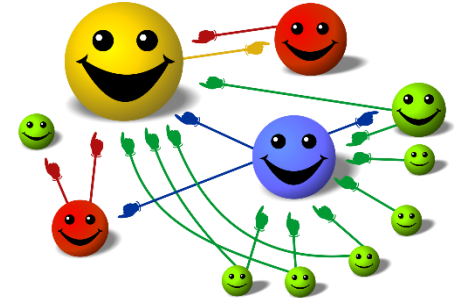
<http://www.google.com/about/datacenters>

Infraestructuras: servidores

- ♦ Cientos de miles de servidores en *data centers* distribuidos por el mundo
- ♦ Servidores Web
 - Recepción de consultas, coordinación de su envío a los servidores de índice
 - Fusión de resultados
 - Obtención de snippets del servidor de documentos, sugerencias de reformulación de los servidores de spelling, anuncios del servidor de publicidad
 - Generación de HTML con resultados
- ♦ Servidores de índices
 - Reciben consultas y devuelven listas de docIDs con score
- ♦ Servidores de documentos
 - Devuelven snippets y documentos completos
- ♦ Servidores de recolección de datos
 - Crawling permanente de la Web, actualización del índice, cómputo de PageRank
- ♦ Servidores de spelling
- ♦ Servidores de publicidad (AdWords, AdSense)

Ránking basado en enlaces: PageRank





- ♦ Aprovechar la estructura de links para extraer indicios de “importancia” genérica de las páginas Web
 - Popularidad, autoridad, calidad...
- ♦ Independientemente de la consulta
 - Después se combinará este criterio con los scores por consulta
- ♦ PageRank: L. Page & S. Brin (T. Winograd) 1995-1998
 - Se suele atribuir a PageRank la efectividad de Google
 - Posteriormente más aplicaciones (PageRank y otros random walks): grafos de palabras, documentos, consultas, clickthrough, tags sociales, redes sociales... (p.e. Twitter WTF)
- ♦ Breve historia
 - 1995 Winograd propone tema de tesis: análisis de la estructura del grafo de la Web; ideas basadas en análisis bibliométrico de citas
 - 1996 Definición de PageRank; arranca un primer crawler; surge la idea de aplicar PageRank a la búsqueda Web
 - 1998 Google se registra como empresa; 60 millones de páginas indexadas (en un garaje); primera financiación
 - 2000 Venta de anuncios asociados a palabras
 - 2003 Googleplex en Mountain View
 - 2004 Google sale a bolsa
 - Hoy Brin & Page, 12º y 13º Forbes



PageRank: principio general

- ◆ Los links entrantes son un indicio de importancia
- ◆ Tanto más si el link procede de una página importante
- ◆ Pero tenemos en cuenta también si la página de origen es muy “pródiga” con los links

Formulación de PageRank

- ♦ Formulación intuitiva, motivación heurística 
- ♦ Interpretación y formalización probabilística 
- ♦ Forma matricial
- ♦ Cómputo de PageRank
 - Autovector principal con $\lambda = 1$
 - Solución algebraica del sistema de ecuaciones (inversión de la matriz del sistema) 
 - Método de potencias (Jacobi, Gauss-Seidel...) → iteración simple
 - Ejemplo de algoritmo sencillo 
- ♦ Teoremas de álgebra y probabilidad aseguran la convergencia

PageRank

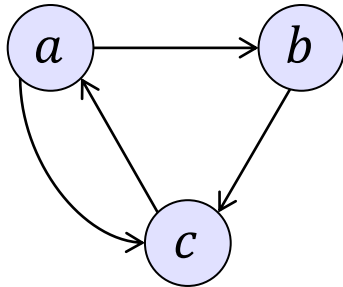
$$P(d_j) = \frac{r}{N} + (1 - r) \sum_{d_i \rightarrow d_j} \frac{P(d_i)}{\#out(d_i)}$$

$$r \in (0,1)$$

PageRank: cómputo iterativo

- ♦ Dar un valor inicial a todas las páginas (p.e. $1/N$)
- ♦ Actualizar el valor de cada página aplicando la fórmula con los valores actuales de las páginas que le apuntan
- ♦ Repetir el paso anterior hasta que se cumpla una condición de convergencia
 - P.e. nº fijo de iteraciones, ~ 50 suele ser suficiente
 - O diferencia de valores de una iteración a la siguiente menor que un umbral
- ♦ Variante: usar progresivamente los valores ya actualizados en el cálculo de las páginas restantes de la misma iteración
 - Acelera la convergencia
- ♦ Importante: tratar los nodos sumidero...

Ejemplo



$$P(a) = r/3 + (1 - r)P(c)$$

$$P(b) = r/3 + (1 - r)P(a)/2$$

$$P(c) = r/3 + (1 - r)(P(a)/2 + P(b))$$

P.e. con $r = 0.5$

- Resolviendo el sistema de ecuaciones $\rightarrow \begin{cases} P(a) = 14/39 \\ P(b) = 10/39 \\ P(c) = 15/39 \end{cases}$
- En general no es viable resolver simbólicamente un sistema de ecuaciones con millones de variables \Rightarrow Cómputo iterativo (solución numérica)

	$k = 1$	$k = 2$	$k = 3$	$k = 4$	$k = 5$	$k = 6$	$k = 7$	$k = 8$	$k = 9$	\dots
$P(a)$	0.3333	0.3333	0.3750	0.3542	0.3594	0.3594	0.3587	0.3590	0.3590	\dots
$P(b)$	0.3333	0.2500	0.2500	0.2604	0.2552	0.2565	0.2565	0.2563	0.2564	\dots
$P(c)$	0.3333	0.4167	0.3750	0.3854	0.3854	0.3841	0.3848	0.3846	0.3846	\dots

PageRank: algoritmo simple

PageRank (*links*)

for $k \leftarrow 1$ to $|links|$ // Compute #outlinks of all nodes

$out[links[k].from] \leftarrow out[links[k].from] + 1$

for $i \leftarrow 1$ to N do // Initial values

$P[i] \leftarrow 1/N$ // (division by N can be omitted)

while *convergence condition* // Compute PageRank iteratively

for $i \leftarrow 1$ to N do

$P'[i] \leftarrow r/N$ // (division by N can be omitted)

for $k \leftarrow 1$ to $|links|$ do

$i \leftarrow links[k].from$

$j \leftarrow links[k].to$

$P'[j] \leftarrow P'[j] + (1 - r) P[i] / out[i]$

for $i \leftarrow 1$ to N do

$P[i] \leftarrow P'[i]$ // To handle sinks, add $(1 - \sum_i P'[i]) / N$

PageRank: interpretación probabilística

- ♦ Navegante aleatorio
 - Empieza en una página al azar
 - Con probabilidad $1 - r$ escoge un enlace saliente al azar y lo atraviesa
 - Con probabilidad r escribe directamente la URL de una página al azar
 - Repite este comportamiento indefinidamente
- ♦ En un instante dado, ¿cuál es la probabilidad $P(d)$ de que este usuario se encuentre en una página d ?
 - $P(d) \equiv \text{PageRank de } d$
- ♦ El escenario describe un proceso estocástico que corresponde a una cadena de Markov: **random walk**
 - Las páginas son estados, el paso de una a otra son transiciones
 - La probabilidad de aterrizar en una página sólo depende de la página anterior
 - La probabilidad de transición de una página a otra se puede calcular
- ♦ Converge a una probabilidad estacionaria
- ♦ Comprobar que resulta la fórmula original...
 - Revela la necesidad de tratamiento de los nodos sumidero

Derivación probabilística

$$p(d_j|t) = \sum_i p(d_j|d_i, t-1)p(d_i|t-1) = \sum_i p(d_j|d_i)p(d_i|t-1)$$

$$p(d_j|d_i) = p(d_j|d_i, click) p(click|d_i) + p(d_j|d_i, teleport) p(teleport|d_i)$$

$$p(click|d_i) = \begin{cases} 1-r & \text{si } \#out(d_i) > 0 \\ 0 & \text{si } \#out(d_i) = 0 \end{cases} \quad p(teleport|d_i) = \begin{cases} r & \text{si } \#out(d_i) > 0 \\ 1 & \text{si } \#out(d_i) = 0 \end{cases}$$

$$p(d_j|d_i, teleport) = p(d_j|teleport) = \frac{1}{N} \quad // \text{ Probabilidad uniforme teleport}$$

$$p(d_j|d_i, click) = \begin{cases} \frac{1}{\#out(d_i)} & \text{si } d_i \rightarrow d_j \\ 0 & \text{en otro caso} \end{cases} \quad // \text{ Probabilidad uniforme entre enlaces}$$

Finalmente, observar que substituyendo se obtiene la fórmula de PageRank donde a los nodos sumidero se les añade un enlace ficticio a todas las demás páginas

Ejemplos con alto PageRank

- ◆ Google no publica esta información
- ◆ Sondeos externos, ver por ejemplo:
 - <https://www.alexacom/topsites>
 - <https://mozcom/top500>
- ◆ Wordpress, Facebook, Twitter, Google, Youtube, Instagram, LinkedIn, Wikipedia, Amazon, Baidu...

Antecedentes en bibliometría

- ◆ Estudio de la estructura de citas entre documentos escritos
 - Medidas de impacto/importancia de documentos y revistas científicas
- ◆ Factor de impacto de Garfield (1955)
 - El impacto de una revista en un año es el número medio de citas entrantes por artículo de los dos años anteriores
 - No tiene en cuenta el factor de impacto del origen las citas
 - Utilizado en el JCR del Institute for Scientific Information (ISI)
- ◆ Medida de Pinski y Narin (1976) + Geller (1978)
 - Las citas de revistas influyentes transmiten más impacto, repartido por igual entre cada cita
 - Equivalente a PageRank con $r = 0$

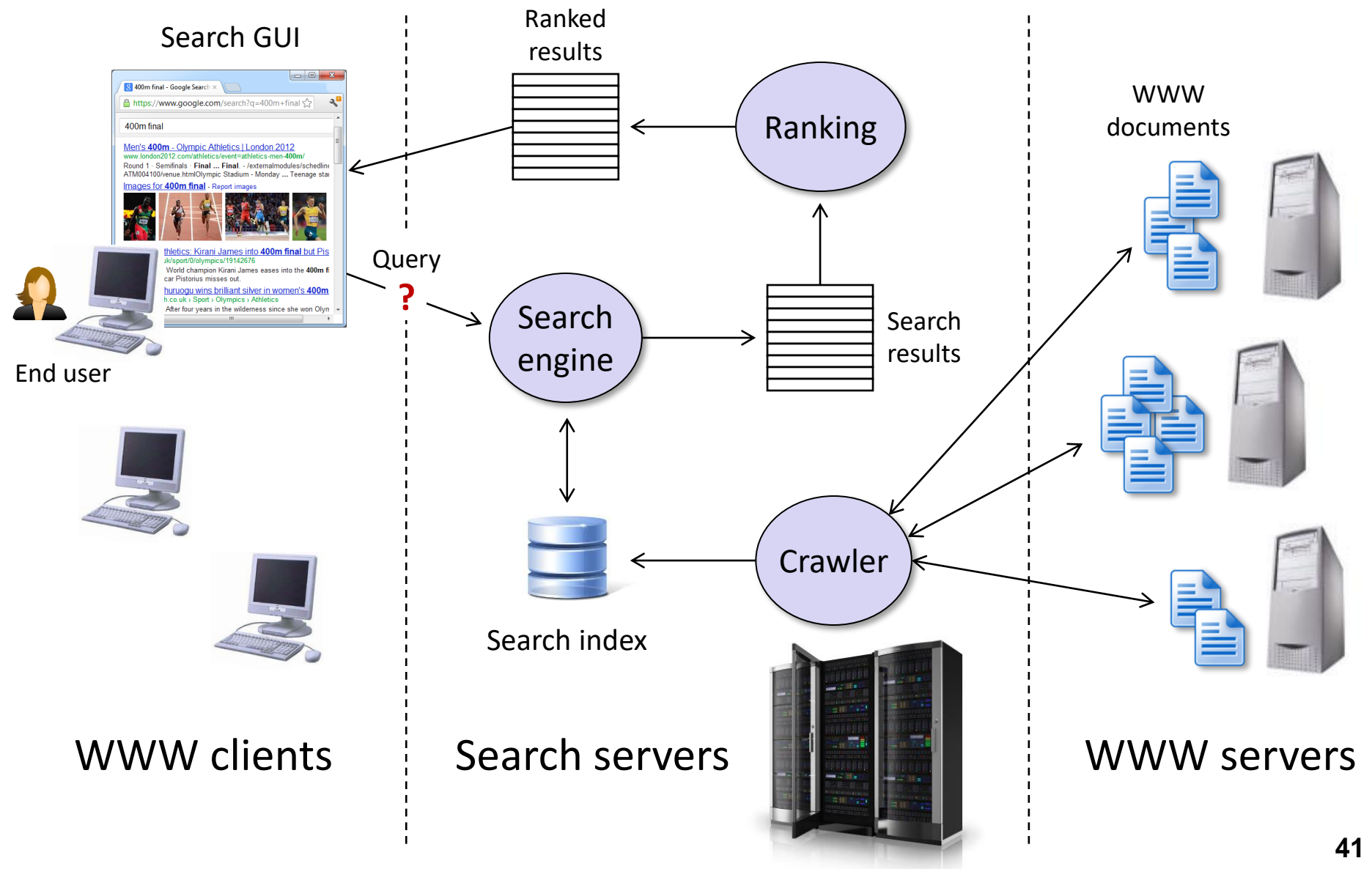
Integración de PageRank en un sistema de búsqueda

- ♦ Elaboraciones monótonas del valor, por ejemplo:
 - No es necesario dividir por N (evita $P(d) \ll 1$ pues $\sum_d P(d) = 1$)
 - $\log P(d)$ modera la distribución tendente a power law
- ♦ Optimización de la convergencia
- ♦ Detectar spam (link farms, etc.)
- ♦ Particularizar el vector de teleportación (p.e. personalización, p.e. en Twitter WTF) y las probabilidades de transición
 - Muchas otras elaboraciones...
- ☞ Combinar $P(d)$ con $\text{sim}(q, d)$
 - Los tres ingredientes principales de Google: contenido + links + RankBrain

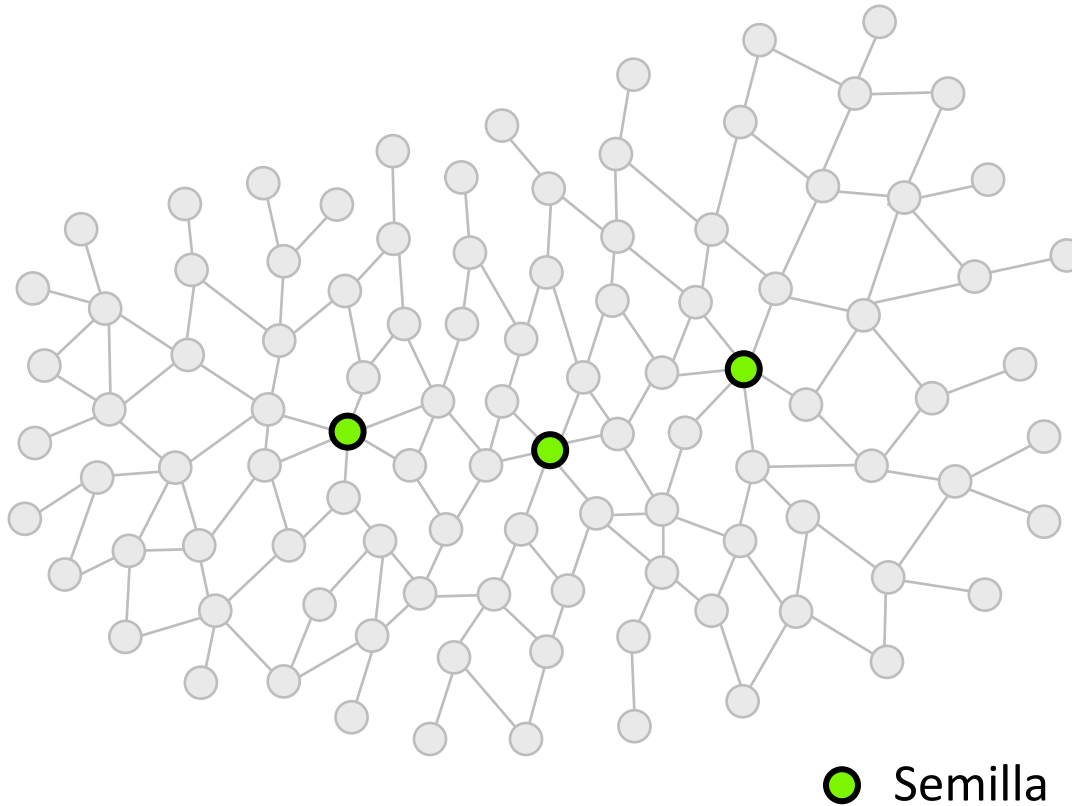
Crawling

- ◆ Crawler, a.k.a. spider, bot, ant, scutter...
- ◆ Componente fundamental de un buscador Web
- ◆ Construcción del índice de búsqueda
- ◆ En una colección distribuida en nodos (servidores) Web
- ◆ Se desconoce de antemano la extensión de la colección y la ubicación de los documentos
- ◆ Explorar la Web es parte del proceso
- ◆ La colección es altamente dinámica
 - Se necesita un crawling continuo

Crawling, indexación y consulta

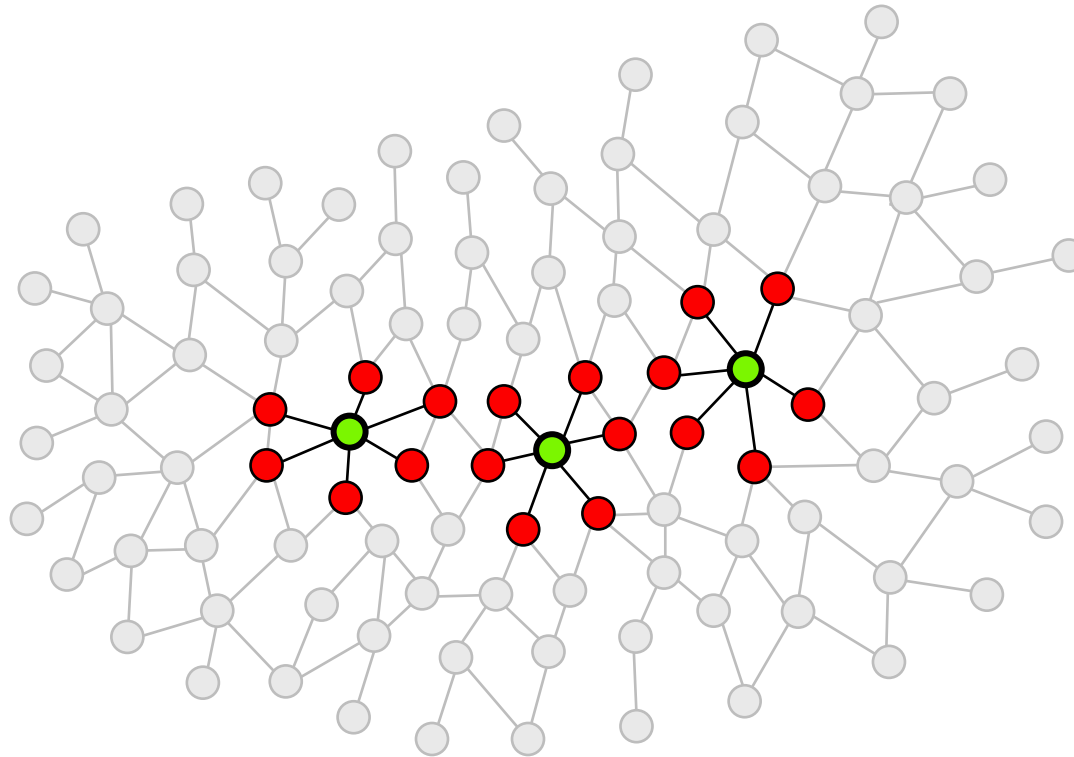


Crawling: exploración de la Web



● Web no descubierta aún

Crawling: exploración de la Web

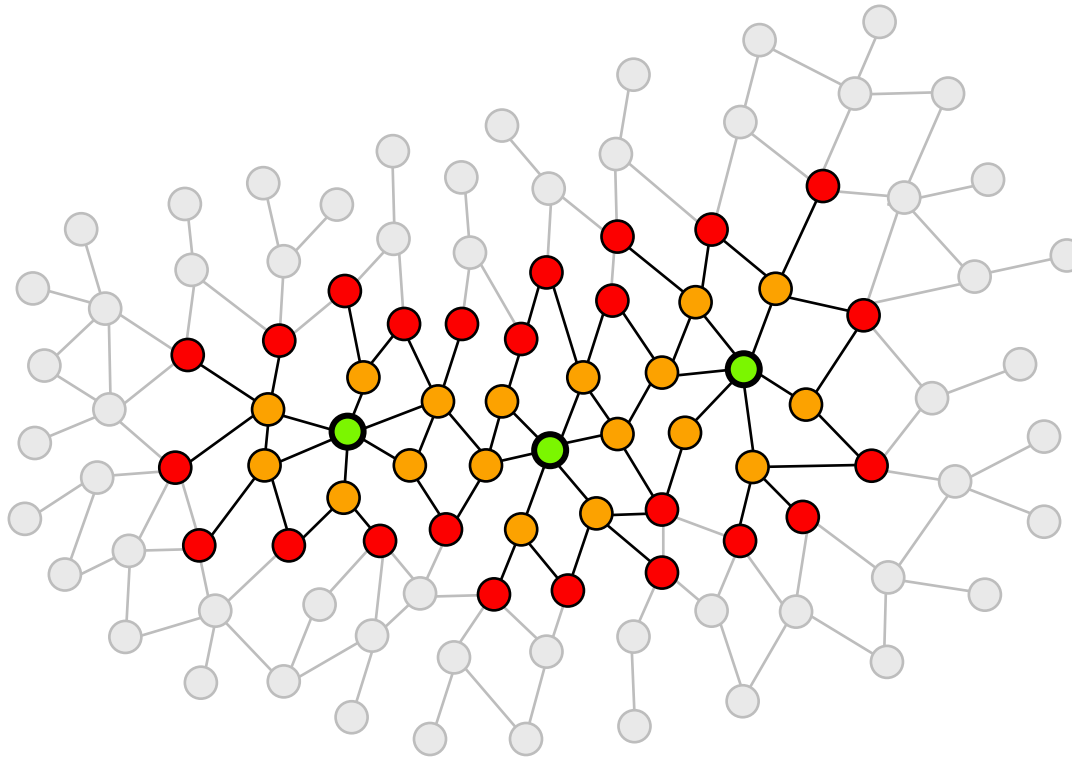


● Semilla

● Frontera de crawling

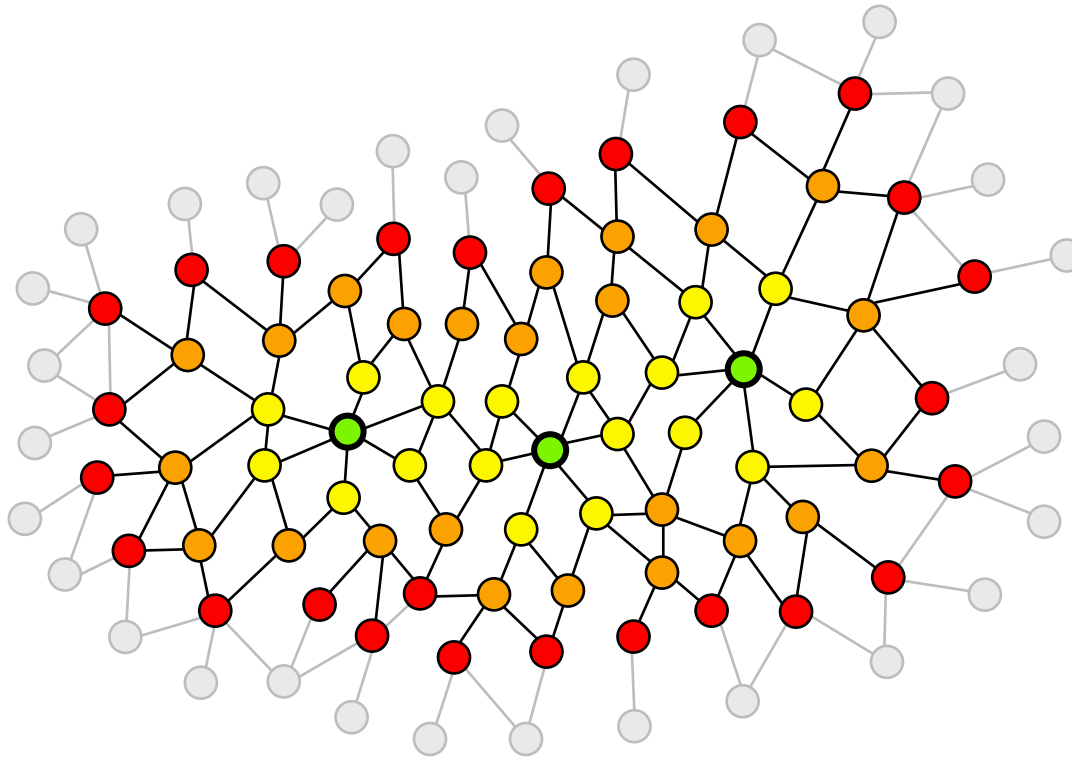
● Web no descubierta aún

Crawling: exploración de la Web



- Semilla
- Web indexada
- Frontera de crawling
- Web no descubierta aún

Crawling: cola de prioridad



Cola de prioridad

● Semilla

● Web indexada

● Frontera de crawling

● Web no descubierta aún

Crawling: pasos

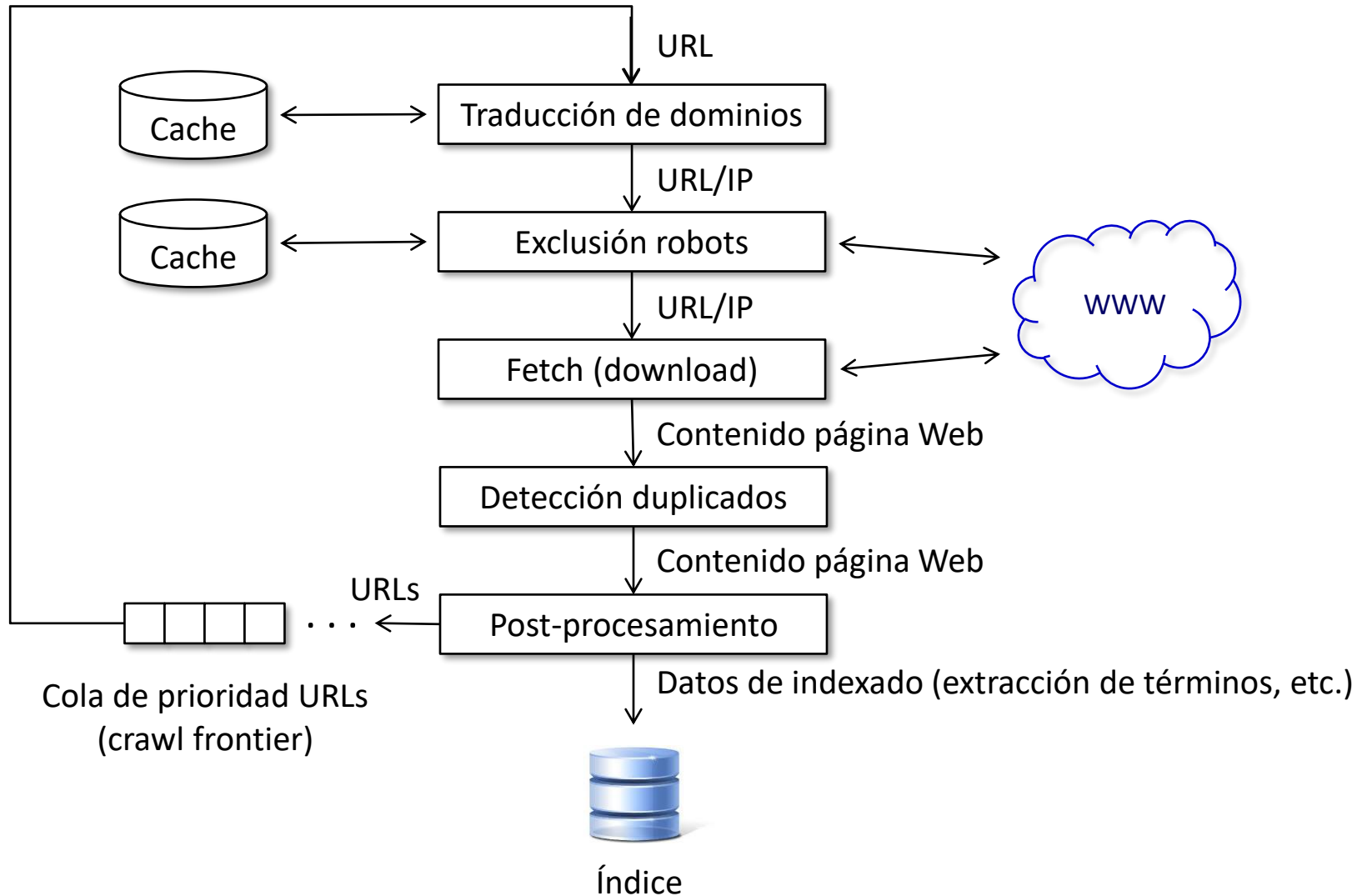
- ◆ Recorrer la Web

- Partir de un conjunto semilla de URLs iniciales
- Variaciones de BFS (cola de prioridad en lugar de simple)

- ◆ Para cada URL

- Descargar el contenido de la página del servidor Web
- Extraer términos de indexado
- Extraer links y recorrerlos (añadir URLs a una cola de prioridad)
- Volver a añadir la URL a la cola de prioridad para su futura actualización



Crawling: pasos (cont)



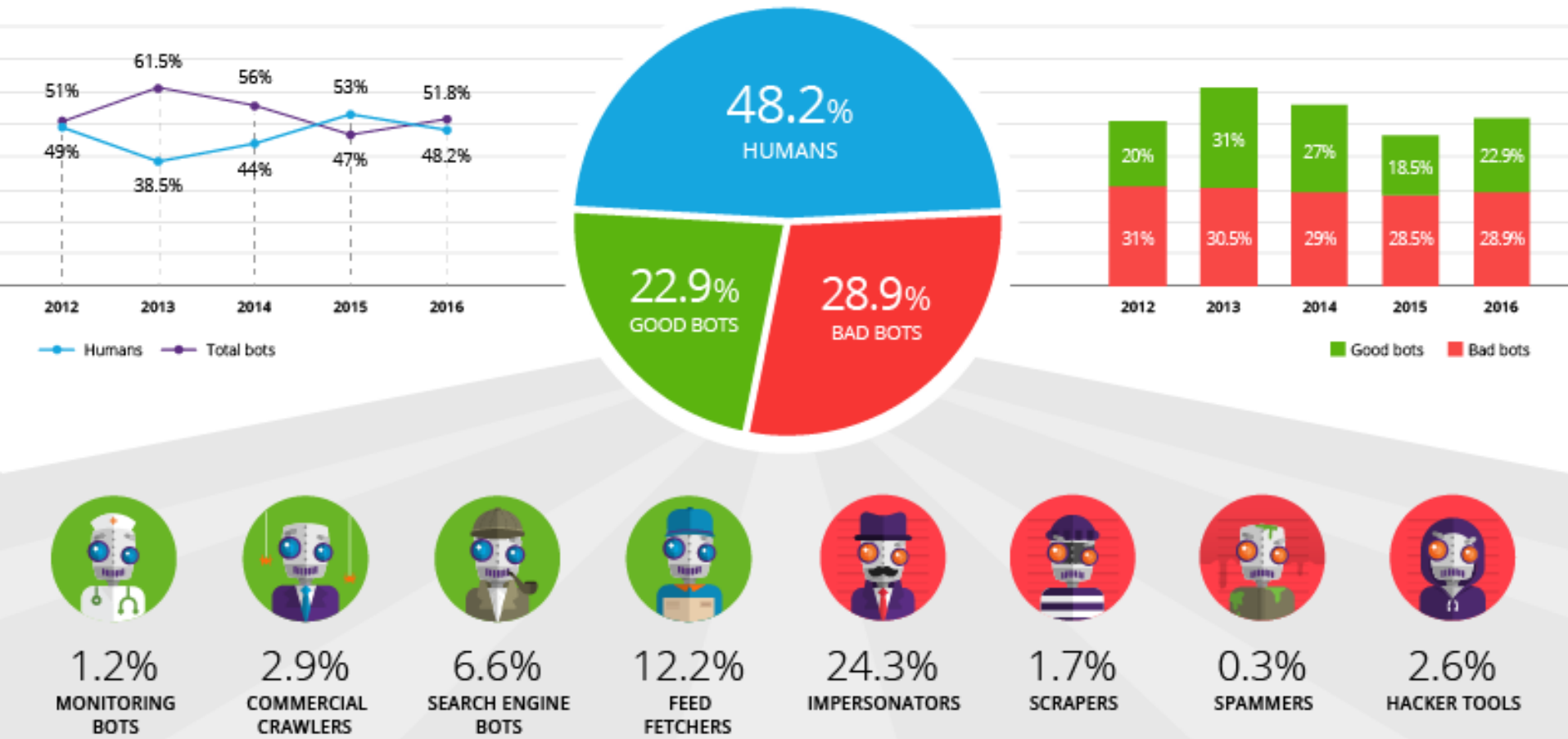
Crawling – operaciones específicas

- ♦ Traducción de dominios
 - La velocidad de un DNS estándar no es suficiente: guardar cache propia
- ♦ Normalización de URLs
 - Mayúsculas en dominio y secuencias de escape, suprimir puerto por defecto, suprimir “.” y “..”, unificar “/” al final de la URL, etc.
- ♦ Aprovechar el texto de los enlaces (términos para el doc apuntado)
 - Se tiene en cuenta con qué términos describen autores externos el contenido de las páginas (p.e. Yahoo! portal no contiene “portal”)
 - Típicamente se ponderan aparte del contenido (p.e. *idf* de “click”, “here”)
 - Vulnerabilidad a ataques de link bombing
- ♦ Si una URL no responde repetidas veces, eliminarla
 - Del índice y de la cola de prioridad
 - Devolver páginas inválidas deteriora la calidad percibida por los usuarios

Propiedades deseables de un crawler

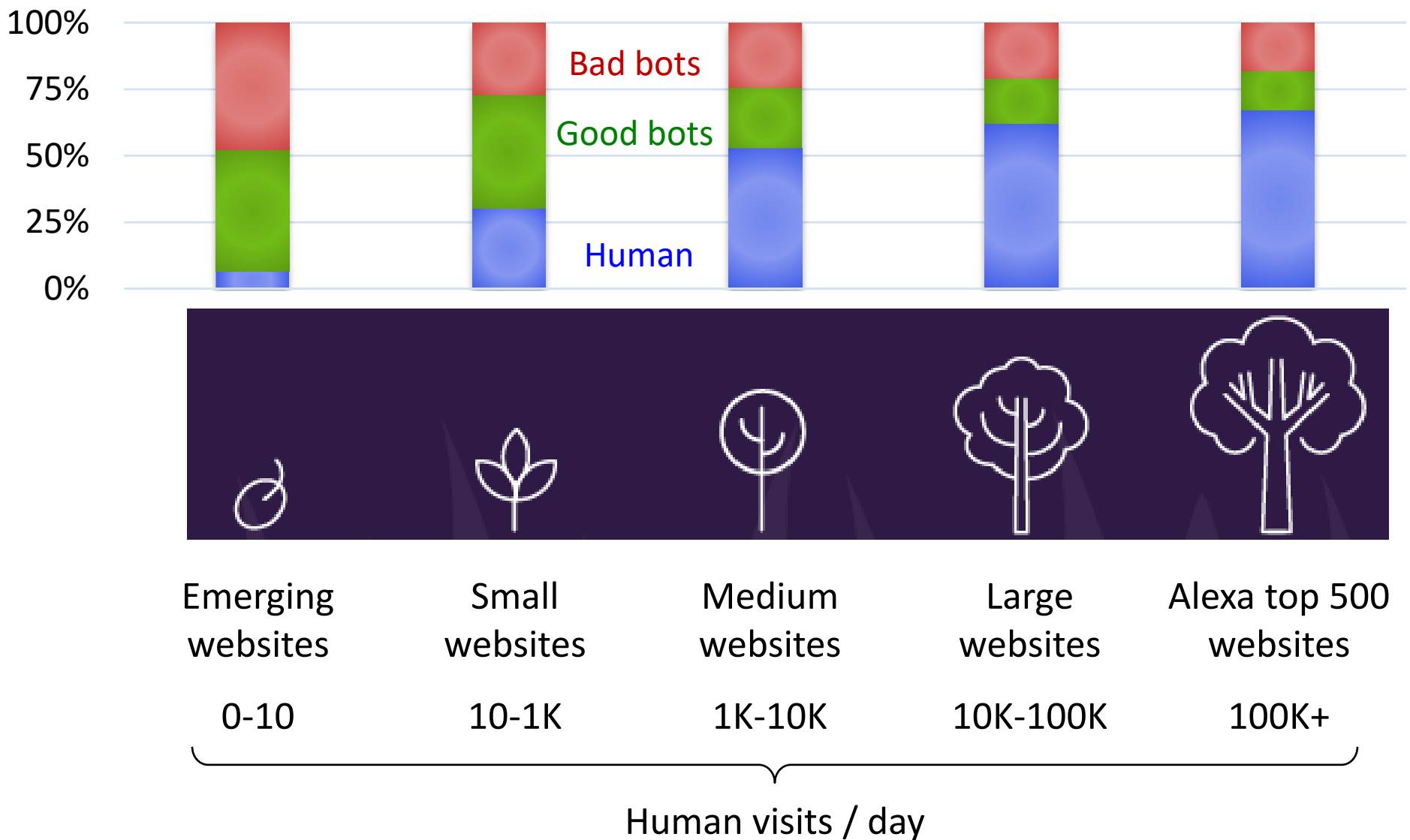
- ♦ Actualidad (freshness) 
- ♦ Calidad: actualización sesgada hacia páginas más útiles
- ♦ Escalabilidad (en particular, distribuibilidad)
- ♦ Cortesía 
- ♦ Robustez
- ♦ Extensibilidad (modularidad)

Cortesía de crawling – tráfico web



<https://www.incapsula.com/blog/bot-traffic-report-2016.html>

Cortesía de crawling – tráfico web



Tráfico de crawling

Ejemplo: un mes de log en ir.ii.uam.es

1 petición cada
± 4 min promedio

Top 15 of 1388 Total User Agents			
#	Hits	User Agent	
1	78582 51.21%	Mozilla/5.0 (Windows; U; Windows NT 5.1; en-US; rv:1.8.1.11) Gecko/20071127 Firefox/2.0.0.11	
2	10781 7.03%	Mozilla/5.0 (compatible; Googlebot/2.1; +http://www.google.com/bot.html)	
3	8268 5.39%	Mozilla/5.0 (compatible; bingbot/2.0; +http://www.bing.com/bingbot.htm)	
4	4301 2.80%	Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKit/537.17 (KHTML, like Gecko) Chrome/24.0.1312.57 Safari/537.17	
5	3540 2.31%	Mozilla/5.0 (Windows NT 6.1; WOW64; rv:18.0) Gecko/20100101 Firefox/18.0	
6	2564 1.67%	Mozilla/5.0 (Windows NT 6.1) AppleWebKit/537.17 (KHTML, like Gecko) Chrome/24.0.1312.57 Safari/537.17	
7	2346 1.53%	Mozilla/5.0 (Windows NT 5.1; rv:18.0) Gecko/20100101 Firefox/18.0	
8	2136 1.39%	Mozilla/5.0 (compatible; Baiduspider/2.0; +http://www.baidu.com/baidu spider)	Firefox version 10 and lower - various robots 35.9%
9	1928 1.26%	Mozilla/5.0 (compatible; MSIE 9.0; Windows NT 6.1; WOW64; Trident/5.0)	Apache-HttpClient 6.7%
10	1559 1.02%	Mozilla/5.0 (Windows NT 6.1; rv:18.0) Gecko/20100101 Firefox/18.0	Googlebot 5.2%
11	1193 0.78%	Mozilla/4.0 (compatible; MSIE 8.0; Windows NT 6.1; WOW64; Trident/4.0)	Sogou web spider 1.3%
12	1183 0.77%	Mozilla/5.0 (compatible; YandexBot/3.0; +http://yandex.com/bot)	bingbot 2.4%
13	1171 0.76%	Mozilla/5.0 (Windows NT 5.1) AppleWebKit/537.17 (KHTML, like Gecko) Chrome/24.0.1312.57 Safari/537.17	python 1.0%
14	968 0.63%	Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKit/537.22 (KHTML, like Gecko) Chrome/24.0.1312.57 Safari/537.17	SemrushBot 0.5%
15	850 0.55%	Mozilla/5.0 (compatible; WBSearchBot/1.1; +http://www.warebot.com)	YandexBot 0.3%
			Baiduspider 0.4%
			BingPreview 0.3%

Ver tiempo entre peticiones consecutivas en un Web log...

Crawling – cortesía con el servidor Web

- ◆ Exclusión de páginas
 - Protocolo de exclusión robots.txt (también se suele guardar en cache):
User-agent, Allow, Disallow
 - `<meta name="robots" content="noindex,nofollow">`
 - `` (muy utilizado automáticamente en blogs y wikis)
- ◆ Moderar el nº de peticiones por minuto
 - P.e. una cada 1-60s, habitualmente > 20s en promedio
 - robots.txt → Crawl-delay
 - Aun así, los crawlers son los mayores consumidores de ancho de banda en Internet
- ◆ Autoidentificarse con el parámetro User-agent en la petición http
 - P.e. Googlebot, Bingbot, Yahoo! Slurp, etc.

Cola de prioridad: actualidad y calidad

- ♦ La gestión de prioridad es el otro problema fundamental a resolver
- ♦ Semillas
 - Portales importantes, portales de noticias, ODP, etc.
- ♦ Estrategia de actualización (avance frontera de crawl)
 - Tiempo de permanencia en la cola
 - Frecuencia y tipo de cambios de las páginas
 - Impacto de los cambios en los rankings de búsqueda
- ♦ Impacto en los rankings
 - Nº de veces que la página aparece en resultados de búsqueda
 - P.e. PageRank, o mejor, frecuencia de clicks de las URLs en un log
- ♦ Tasas de actualización del contenido
 - P.e. portales de noticias se pueden reindexar cada hora o más veces
 - Otras páginas más estables cada varias semanas
 - Distinguir cambios sin importancia (p.e. ads, “quote of the day”, etc.)

Actualidad y escalabilidad: paralelización

- ♦ Imposible realizar un recorrido en proceso secuencial
 - Latencia y capacidad de respuesta de los servidores Web recorridos
 - Para indexar la Web en un mes se necesitan procesar muchos GB/s
- ♦ Cómo organizar la paralelización es uno de los problemas fundamentales a resolver en el desarrollo de un crawler
 - P.e. un hilo por URL, o batches de URLs
 - Amplio número de servidores, cada uno se ocupa de una porción de la Web (reparto por direcciones IP, dominios, etc.)
 - La cola de prioridad puede ser centralizada o distribuida también

Crawling – páginas dinámicas

- ♦ Muchas no se pueden indexar
 - P.e. toman input del usuario (front-end de aplicaciones), acceso vía enlaces creados con JavaScript, redirects, protegidas por password, etc.
- ♦ Otras sí
 - Existe un camino de enlaces visible para el crawler: equivale a una página estática
 - No existe camino de enlaces: archivos sitemap (p.e. catálogos de tiendas online, etc.)
- ♦ Archivos sitemap (ver <http://www.sitemaps.org>)
 - Propuesto por Google en 2005, secundado poco después por Yahoo, MSN, Ask, IBM...
 - Contienen un listado de URLs a indexar, con detalles de prioridad (relativa), periodicidad de actualización, fecha de última modificación, etc.
 - La URL de ubicación del archivo sitemap se indica en robots.txt
 - Se suelen admitir máx 50.000 URLs / 10MB por sitemap, permitiendo un archivo índice de sitemaps con los mismos máximos (generalmente comprimido con gzip)
 - El servidor vuelca sus URLs dinámicas (p.e. generan consultas a una BD) a un listado sitemap en XML, el crawler las incluye en la cola de URLs
- ♦ También es posible enviar una URL manualmente para solicitar indexación

Web spam

- ♦ Manipular favorablemente la posición de ránking de una página
 - Para determinadas palabras clave: multiplicar artificialmente la frecuencia de las palabras en el documento (p.e. invisibles al usuario)
 - Independiente de consulta: proliferar enlaces artificiales (link farms, intercambio de links, etc.) que promocionen engañosamente el PageRank del documento
 - Simular clicks con programas
- ♦ Hacerse indexar por palabras diferentes a las que realmente contiene una página
 - Cloaking: el servidor Web entrega contenido diferente según el agente sea un usuario (contenido real) o un crawler (doorway page)
- ♦ Sabotear páginas de terceros
 - Manipulando los términos que las indexan: enlaces masivos con texto peyorativo (link bombing), por programa o consigna entre usuarios
 - Perjudicando su posición de ránking: simulando técnicas fraudulentas (p.e. link farming) por parte de la página (Google bowling)

Link bombing



Web

5 results stored on your computer - [Hide](#) - [About](#)



[About Blether](#) - a gentle term for a **liar** or fibber, someone who is telling
[BBC NEWS | Programmes | F..](#) - of being a **liar**, a dictator, of robbing

[Tony Blair - Biography](#)

Read the full biography of Prime Minister Tony Blair.

[www.pm.gov.uk/output/Page4.asp](#) - 12k - [Cached](#) - [Similar pages](#) - [Remove result](#)

[Web](#) [Images](#) [Maps](#) [News](#) [Shopping](#) [Gmail](#) [more ▼](#)



dangerous cult

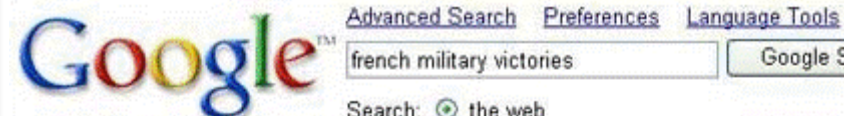
Search

Web

[Scientology - Church of Scientology Official Site](#)

Living in a **Dangerous** Environment · Drug and Alcohol Problems · Personalities, Emot
and How to Deal with Others ...

[www.scientology.org/](#) - 74k - [Cached](#) - [Similar pages](#)



Search: ☒ the web

[Web](#) [Images](#) [Groups](#) [Directory](#) [News](#)

Did you mean: [french military defeats](#)

No standard web pages containing all your search terms were found.

Your search - **french military victories** - did not match any documents.

Suggestions:

- Make sure all words are spelled correctly.
- Try different keywords.
- Try more general keywords.
- Try fewer keywords.

Web spam (cont)

◆ Spider traps

- El servidor genera un nº ilimitado de URLs en las que el crawler queda atrapado
- A menudo no intencionado (p.e. links en calendar online)
- Solución: poner un límite a la longitud de las URLs, al nº de páginas a indexar en un mismo sitio Web, a la profundidad de crawl... no hay una solución perfecta

◆ Los buscadores necesitan sofisticadas medidas anti-spam

- La detección de spam es un área amplia de la recuperación de información: adversarial information retrieval
- Penalización y bloqueo de las manipulaciones detectadas
- Evolución mutua continua de técnicas anti-spam y spam
- Indicaciones técnicas SEO toleradas / no toleradas
- P.e. Google menciona un $\sim 0.22\%$ de dominios marcados a eliminar por spam