

# Técnicas para la recuperación de información en la Web

Sistemas Inteligentes  
de Acceso a la Información

José María Gómez Hidalgo

<http://www.esp.uem.es/jmgomez/sinai/>

Sistemas Inteligentes de Acceso a la Información – José María Gómez Hidalgo 1  
<http://www.esp.uem.es/~jmgomez/sinai/>

## Índice

- ***Introducción***
- Características de la Web
- Técnicas de búsqueda
  - Motores de búsqueda
  - Directorios
  - Navegación
- Estudios de usuarios
- Resumen

Sistemas Inteligentes de Acceso a la Información – José María Gómez Hidalgo 2  
<http://www.esp.uem.es/~jmgomez/sinai/>

## Introducción

- **World Wide Web**
  - Sistema de hipertexto que opera sobre Internet, usado para servir páginas Web y transmitir archivos
- **Construido sobre varios elementos**
  - La idea de hipertexto (navegación conceptual)
  - Los identificadores de recursos (URLs, URNs)
  - Un modelo de computación cliente-servidor con un protocolo específico (HTTP)
  - Un lenguaje de marcas (HTML)

Sistemas Inteligentes de Acceso a la Información – José María Gómez Hidalgo 3  
<http://www.esp.uem.es/~jmgomez/sinai/>

## Introducción

- **Surgida en los 80 (Tim Berners-Lee, CERN)**
- **Abierta al público en 1993**
  - Inmediata aceptación y popularización
  - Simplifica sistemas previos (Gopher)
  - Crecimiento exponencial
- **Hoy**
  - Quizá el sistema de información más popular y complejo del mundo
  - Impone numerosas dificultades para la búsqueda

Sistemas Inteligentes de Acceso a la Información – José María Gómez Hidalgo 4  
<http://www.esp.uem.es/~jmgomez/sinai/>

## Introducción

- Algunos retos
  - Sistema distribuido
  - Alto porcentaje de volatilidad
  - Enorme tamaño
  - Datos sin estructura y redundantes
  - Calidad potencialmente deficiente de los datos
    - Tipos, formatos

Sistemas Inteligentes de Acceso a la Información – José María Gómez Hidalgo 5  
<http://www.esp.uem.es/~jimgomez/sinai/>

## Índice

- **Introducción**
- **Características de la Web**
- Técnicas de búsqueda
  - Motores de búsqueda
  - Directorios
  - Navegación
- Estudios de usuarios
- Resumen

Sistemas Inteligentes de Acceso a la Información – José María Gómez Hidalgo 6  
<http://www.esp.uem.es/~jimgomez/sinai/>

## Características de la Web

- Múltiples aspectos a considerar
  - Número de servidores, sitios, páginas y archivos
  - Tipo, tamaño y formato de los archivos
  - Contenidos de las páginas
  - Sitios más importantes
  - Distribución de idiomas
  - Estructura física
  - Etc.

Sistemas Inteligentes de Acceso a la Información – José María Gómez Hidalgo 7  
<http://www.esp.uem.es/~jimgomez/sinai/>

## Características de la Web

- Características
  - Difícil de estimar en un momento fijo (no hay una foto fija de la Web)
  - Múltiples estudios, metodologías y estimaciones
  - Los disponibles en [Baeza-Yates], obsoletos (<98)
- [Bharat98]
  - Metodología: Léxico de 400K palabras (Yahoo) => 20K consultas aleatorias => 4 motores de búsqueda => solapamiento y cobertura
  - Resultados (Julio 1998): 350M de páginas

Sistemas Inteligentes de Acceso a la Información – José María Gómez Hidalgo 8  
<http://www.esp.uem.es/~jimgomez/sinai/>

## Características de la Web

- Varios (1995)
  - Muestras de páginas (11M, 2.6M)
  - Tipo y popularidad
    - HTML > GIF, JPEG > ASCII > PS
  - Páginas web (HTML)
    - 5Kb en media, 2Kb mediana
    - 1-2 imágenes por página, 14Kb por imagen
    - 5-15 enlaces salientes por página, >8 en media
    - Mayoría de enlaces locales
    - <1 enlace entrante por página

Sistemas Inteligentes de Acceso a la Información – José María Gómez Hidalgo 9  
<http://www.esp.uem.es/~jmgomez/sinai/>

## Características de la Web

- Varios (1995)
  - Sitios más populares: Microsoft, Netscape, Yahoo!, universidades EEUU
- Idiomas (1998)
  - Web: Inglés (71%), Alemán (7%), Japonés (4%), Francés (3%), Español (3%), Portugués (2%), Italiano (1%)
  - Hablantes: Inglés (450M), Alemán (118M), Japonés (126M), Francés (122M), Español (266M), Portugués (175M), Italiano (63M)

Sistemas Inteligentes de Acceso a la Información – José María Gómez Hidalgo 10  
<http://www.esp.uem.es/~jmgomez/sinai/>

## Características de la Web

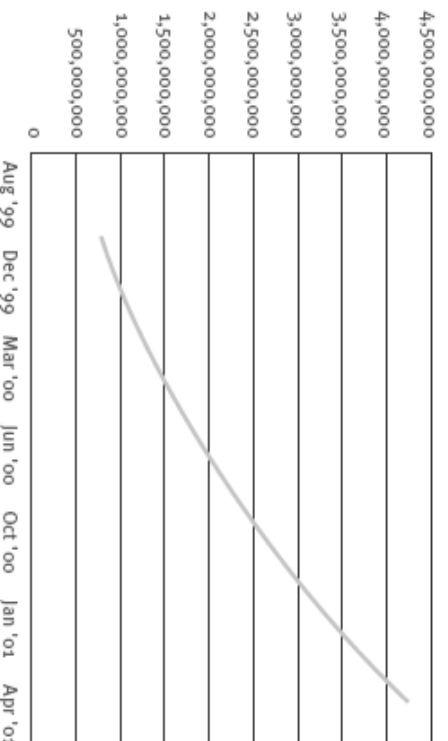
- [Moore00]
  - Metodología: software de análisis propietario (NetSapien) => monitorización de 350M enlaces X 4 meses
  - Resultados (Julio 2000): 2100M páginas, 7M páginas agregadas X día
  - Tasa de crecimiento cuadrático
    - 3000M (Octubre 2000), 4000M (Febrero 2001)

Sistemas Inteligentes de Acceso a la Información – José María Gómez Hidalgo  
<http://www.esp.uem.es/~jmgomez/sinai/>

11

## Características de la Web

- Número de páginas y archivos [Moore00]



Sistemas Inteligentes de Acceso a la Información – José María Gómez Hidalgo  
<http://www.esp.uem.es/~jmgomez/sinai/>

12

## Características de la Web

- [Moore00]
  - Páginas Web (HTML)
    - 10Kb por página
    - 23 enlaces internos por página (mediana 4)
    - 5,6 enlaces externos por página (mediana 1)
    - 14,38 imágenes por página (mediana 1)
    - 84,7% páginas en EEUU, 15,37% resto

Sistemas Inteligentes de Acceso a la Información – José María Gómez Hidalgo  
<http://www.esp.uem.es/~jimgomez/sinai/>

13

## Características de la Web

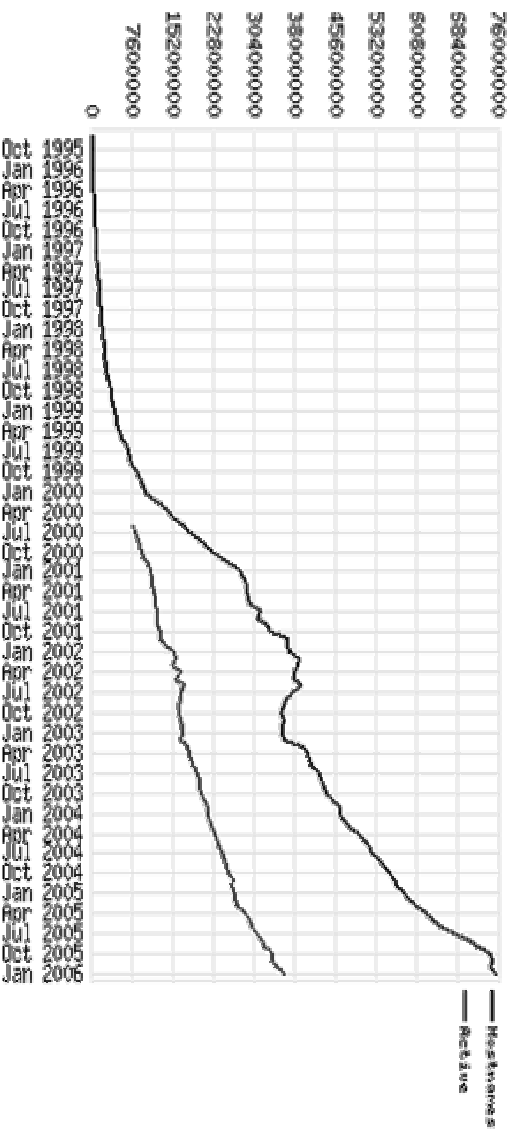
- [Netcraft] Web Server Survey
  - Metodología: *script* inicialmente (1995) alimentado con nombres de *host* públicos, que descubre nuevos servidores
  - Monitorización permanente 95-hoy
    - Servidores vivos y muertos
    - Fabricantes de servidores web

Sistemas Inteligentes de Acceso a la Información – José María Gómez Hidalgo  
<http://www.esp.uem.es/~jimgomez/sinai/>

14

## Características de la Web

- [Netcraft] Web Server Survey

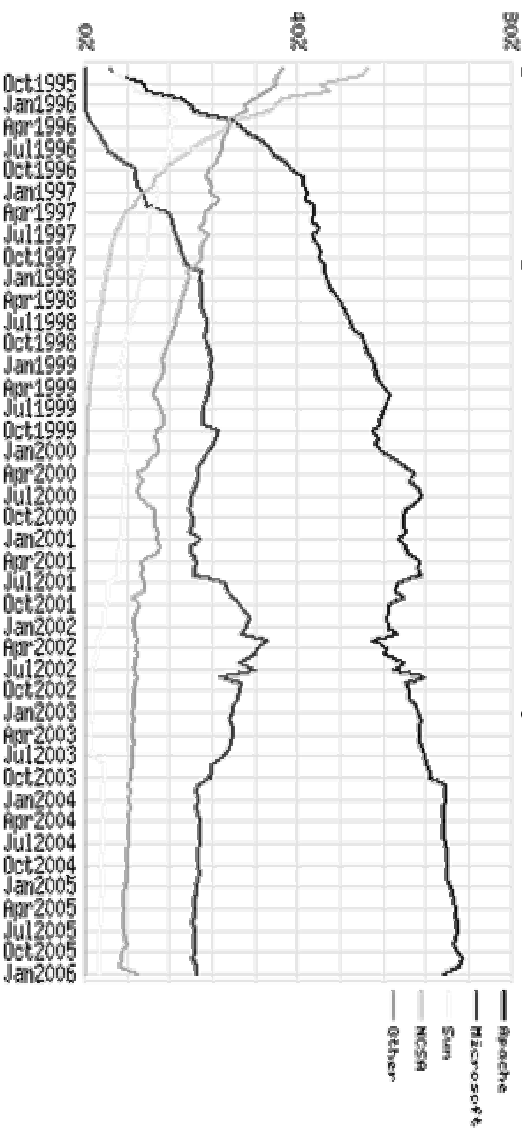


Sistemas Inteligentes de Acceso a la Información – José María Gómez Hidalgo  
<http://www.esp.uem.es/~jmgomez/sinai/>

15

## Características de la Web

- [Netcraft] Web Server Survey

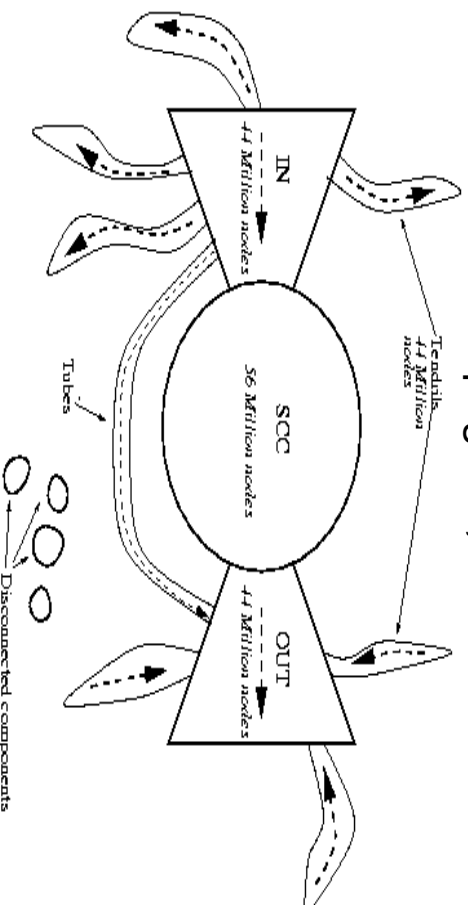


Sistemas Inteligentes de Acceso a la Información – José María Gómez Hidalgo  
<http://www.esp.uem.es/~jmgomez/sinai/>

16

# Características de la Web

- [Broder00] Topología
  - Muestra de 200M páginas, 1500M enlaces



Sistemas Inteligentes de Acceso a la Información – José María Gómez Hidalgo  
<http://www.esp.uem.es/~jmgomez/sinai/> 17

## Índice

- **Introducción**
- **Características de la Web**
- **Técnicas de búsqueda**
  - Motores de búsqueda
  - Directorios
  - Navegación
- **Estudios de usuarios**
- **Resumen**

Sistemas Inteligentes de Acceso a la Información – José María Gómez Hidalgo  
<http://www.esp.uem.es/~jmgomez/sinai/> 18

# Técnicas de búsqueda

- Tres tipos básicos
  - Motores de búsqueda
    - Herramientas de recuperación en sentido clásico
    - Google, Altavista, etc.
  - Directorios
    - Catálogos manuales de sitios
    - Yahoo, Open Directory project, etc.
  - Navegación
    - Usan los hiperenlaces para nuevas funcionalidades
    - Agentes como Alexa o WebGlimpse

Sistemas Inteligentes de Acceso a la Información – José María Gómez Hidalgo  
<http://www.esp.uem.es/~jmgomez/sinai/>

19

# Índice

- **Introducción**
- **Características de la Web**
- **Técnicas de búsqueda**
  - **Motores de búsqueda**
  - Directorios
  - Navegación
- Estudios de usuarios
- Resumen

Sistemas Inteligentes de Acceso a la Información – José María Gómez Hidalgo  
<http://www.esp.uem.es/~jmgomez/sinai/>

20

## Motores de búsqueda

- **Sistemas de recuperación en sentido clásico**
  - Ciclo básico de interacción (consulta => lista de documentos)
  - Pocas funcionalidades adicionales
    - Operadores booleanos, de proximidad, de frase
    - Más como este, traducciones
  - Técnicas tradicionales (ampliadas)
    - Modelo booleano, MEV
    - Tokenización, stoplist, stemming, ranking por similitud

Sistemas Inteligentes de Acceso a la Información – José María Gómez Hidalgo

<http://www.esp.uem.es/~jmgomez/sinai/>

21

## Motores de búsqueda

- **Tipos básicos**
  - Generalistas (Google, Altavista)
  - Meta-buscadores (MetaCrawler, DogPile)
  - Verticales o temáticos (ejemplos)
    - Buscadores de buscadores (CompletePlanet, TheInvisibleWeb)
    - Noticias (Feeder, Google News, RocketNews)
    - Trabajo (Monster, InfoJobs)
    - Viajes (Mobilissimo, Viajar, Rumbo)
    - Artículos científicos (Citeaser, Google Scholar)

Sistemas Inteligentes de Acceso a la Información – José María Gómez Hidalgo

<http://www.esp.uem.es/~jmgomez/sinai/>

22

## Motores de búsqueda

- Estructura afectada principalmente por
  - Web distribuida => módulo de descarga de objetos a indexar = crawler, spider, bot
  - Tamaño y volatilidad de los datos + requerimiento de permanencia en servicio => actualización y búsquedas simultáneas
  - Existencia de hiperenlaces + calidad de recursos => ranking incorpora elementos adicionales (popularidad en enlaces, visitas, etc.)

Sistemas Inteligentes de Acceso a la Información – José María Gómez Hidalgo

<http://www.esp.uem.es/~jmgomez/sinai/>

23

## Motores de búsqueda

- Tecnologías usadas propietarias
  - Patentadas, indocumentadas
  - Información en white-papers y algunos artículos (Conferencias WWW)
- Crawler (Robot, Bot)
  - Función = descarga rápida de páginas Web para poblar el índice del buscador
  - Diseñados para la eficiencia y la explotación óptima de recursos (multi-hilo, etc.)

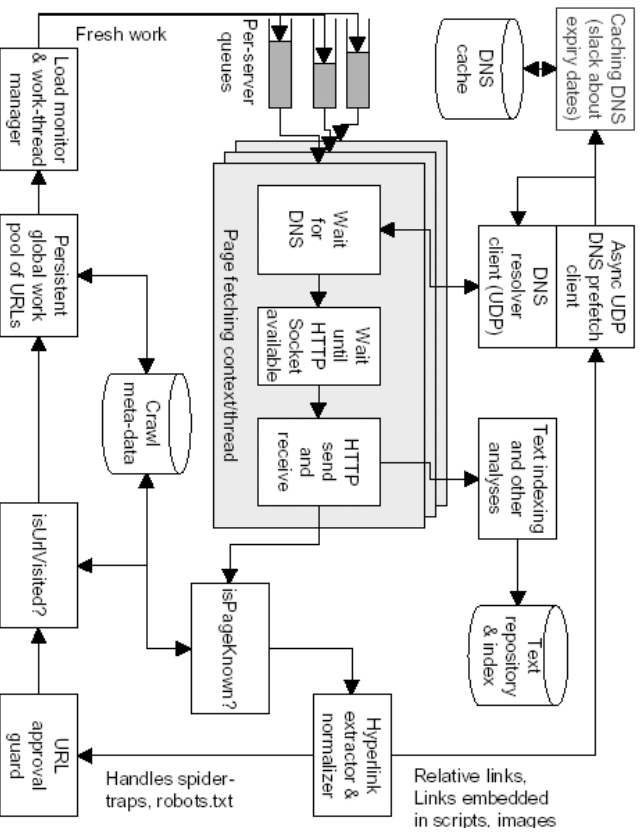
Sistemas Inteligentes de Acceso a la Información – José María Gómez Hidalgo

<http://www.esp.uem.es/~jmgomez/sinai/>

24

## Motores de búsqueda

- Anatomía (Mercator, base de Altavista) 1999



Sistemas Inteligentes de Acceso a la Información – José María Gómez Hidalgo  
<http://www.esp.uem.es/~jmgomez/sinai/>

25

## Motores de búsqueda

- Crawler
  - Separado del indexador por eficiencia
  - Usa caches locales para optimizar ancho banda
  - Usa estructuras *hash* (MD5) para optimizar tiempo
  - Usa módulos específicos de control de balanceo de carga
  - Debe preocuparse de páginas dinámicas, spider-traps, robots.txt, ataques DOS, modificación de páginas, etc.

Sistemas Inteligentes de Acceso a la Información – José María Gómez Hidalgo  
<http://www.esp.uem.es/~jmgomez/sinai/>

26

## Motores de búsqueda

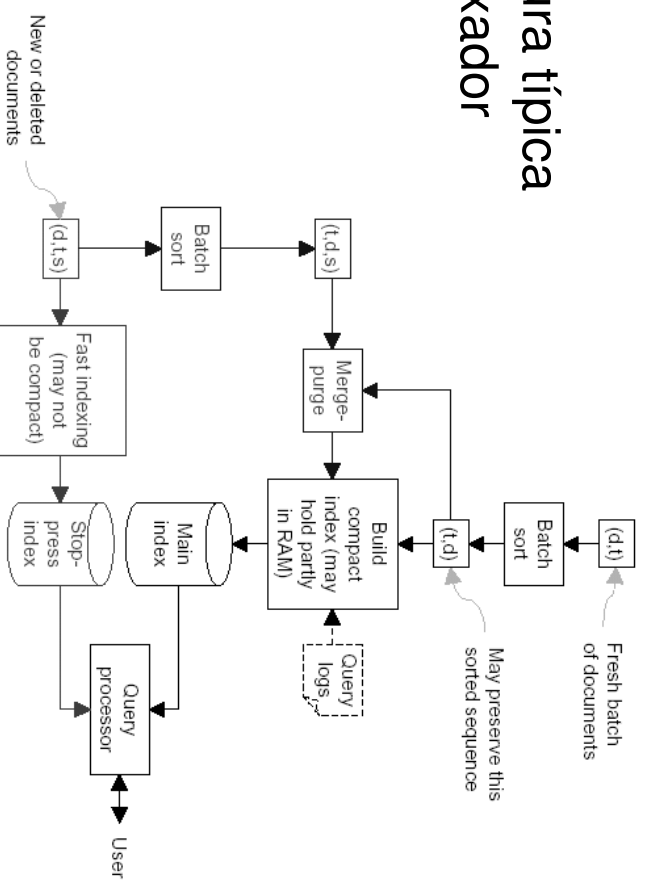
- Indexación
  - Debe realizarse durante la operación del sistema (búsquedas) + colección dinámica (adiciones, eliminaciones) => sus resultados inmediatamente incorporados a un índice parcial paralelo, absorbido periódicamente
  - Usa algoritmos rápidos y distribuidos, y estructuras eficientes para su gestión (e.g. Inversión rápida de índices)

Sistemas Inteligentes de Acceso a la Información – José María Gómez Hidalgo  
<http://www.esp.uem.es/~jmgomez/sinai/>

27

## Motores de búsqueda

- Estructura típica de indexador



Sistemas Inteligentes de Acceso a la Información – José María Gómez Hidalgo  
<http://www.esp.uem.es/~jmgomez/sinai/>

28

## Motores de búsqueda

- Técnicas de ranking por popularidad
  - Popularidad medida en enlaces / visitas
- Popularidad basada en enlaces
  - Técnicas basadas en análisis de citas
    - Calidad e impacto de revistas y artículos científicos
  - Múltiples conceptos: prestigio, centralidad, etc.
- E = matriz adyacencia grafo dirigido de citas
  - $E[i,j] = 1$  si el documento i cita al j, 0 en otro caso
  - Grafo dirigido  $E[i,j] = 1 \iff E[j,i] = 1$

Sistemas Inteligentes de Acceso a la Información – José María Gómez Hidalgo

<http://www.esp.uem.es/~jmgomez/sinai/>

29

## Motores de búsqueda

- Prestigio
  - Prestigio = valor numérico  $p[v]$  asociado a cada nodo  $v \Rightarrow P$  es un vector sobre  $N$  vértices
  - Concepto recursivo
    - $A \rightarrow B, B \rightarrow C \Rightarrow$  prestigio de  $C =$  una función del prestigio de  $B =$  una función del prestigio de  $A = \dots$
  - El prestigio asociado a  $v$  es la suma del prestigio de sus referentes

$$p'[v] = \sum_u E[u,v].p[u]$$

Sistemas Inteligentes de Acceso a la Información – José María Gómez Hidalgo

<http://www.esp.uem.es/~jmgomez/sinai/>

30

## Motores de búsqueda

- Prestigio
  - Proceso iterativo  
 $p \leftarrow E^T p$
  - Converge a punto fijo = autovector principal de E
  - Inicio con  $p = [1, 1, 1, \dots, 1]$
  - Tras cada iteración, normalizar para evitar desbordamiento numérico usando la norma 1

$$\|p\|_1 = \sum_u p[u]$$

Sistemas Inteligentes de Acceso a la Información – José María Gómez Hidalgo 31  
<http://www.esp.uem.es/~jmgomez/sinai/>

## Motores de búsqueda

- Centralidad
  - Determina la influencia / impacto de un nodo
  - Distancia de u a v (longitud del camino mínimo) =  $D[u, v]$
  - Radio de u =  $R(u) = \max_v (D[u, v])$
  - Nodos centrales = los de menor radio = los más influyentes
    - OBS: en caso de citas, invertir los enlaces o caminos (interesan los más citados)

Sistemas Inteligentes de Acceso a la Información – José María Gómez Hidalgo 32  
<http://www.esp.uem.es/~jmgomez/sinai/>

## Motores de búsqueda

- PageRank
  - Similar al prestigio con algunas modificaciones
  - Simula el comportamiento de un humano explorando sin fin la Web sin objetivo definido
    - Selecciona un enlace al azar en cada página
  - $p_0[u]$  = probabilidad inicial de estar en una página
  - Cada página tiene  $N_u$  enlaces salientes, y asumimos enlaces  $[u, v]$  únicos

$$N_u = \sum_v E[u, v] \quad p_1[v] = \sum_{(u,v) \in E} \frac{p_0[u]}{N_u}$$

Sistemas Inteligentes de Acceso a la Información – José María Gómez Hidalgo  
<http://www.esp.uem.es/~jmgomez/sinai/>

33

## Motores de búsqueda

- PageRank
  - Algebraicamente, de computa la matriz  $L$  = normalización de  $E$  con filas en suma 1
$$L[u, v] = \frac{E[u, v]}{\sum_{\beta} E[u, \beta]} = \frac{E[u, v]}{N_u}$$
  - Con  $L$ , el cómputo de  $p_1$  es directo
$$p_1[v] = \sum_u L[u, v] \cdot p_0[u] \quad p_1 = L^T p_0$$
  - Y en general  $p_{i+1} = L^T p_i$

Sistemas Inteligentes de Acceso a la Información – José María Gómez Hidalgo  
<http://www.esp.uem.es/~jmgomez/sinai/>

34

## Motores de búsqueda

- PageRank
  - La ecuación anterior converge al autovector principal de  $L^T =$  solución de  $p=L^T p =$  distribución estacionaria de  $L$ , independiente de  $p_0$
  - Basado en grafo fuertemente conectado, sin ciclos
    - Para evitar ciclos infinitos => en cada nodo, el viajante salta con probabilidad  $d \in [0, 1, 0.2]$  a una página cualquiera, y  $1-d$  a una página vecina

Sistemas Inteligentes de Acceso a la Información – José María Gómez Hidalgo 35  
<http://www.esp.uem.es/~jmgomez/sinai/>

## Motores de búsqueda

- PageRank
  - El modelo ampliado es
$$p_{i+1} = \left( (1-d).L^T + \frac{d}{N}.1_N \right) .p_i$$
  - Se calculan las iteraciones necesarias para llegar a un resultado suficientemente estable
  - Se normaliza periódicamente para evitar desbordamientos numéricos

Sistemas Inteligentes de Acceso a la Información – José María Gómez Hidalgo 36  
<http://www.esp.uem.es/~jmgomez/sinai/>

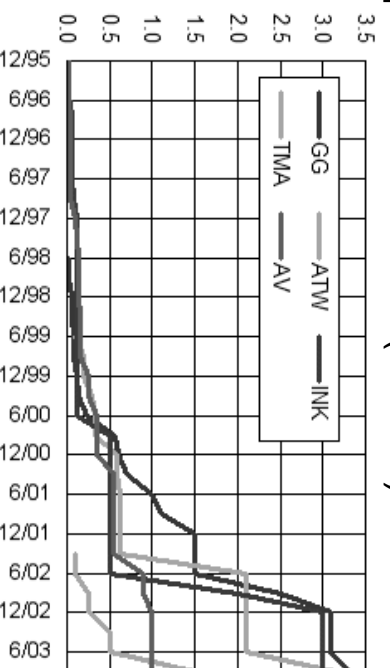
## Motores de búsqueda

- PageRank
  - En Google se combina PageRank+MEV
  - Se suele criticar porque desacopla la calidad del tema de la página
  - Es susceptible de ataques basados en enlaces
    - Incrementar el PageRank artificialmente por medio de enlaces ad-hoc
    - Search Engine Spamming (<http://noseencuentra.com>)
    - Distinto de Search Engine Optimization (campanñas de marketing éticas, <http://searchenginewatch.com>)

Sistemas Inteligentes de Acceso a la Información – José María Gómez Hidalgo 37  
<http://www.esp.uem.es/~jmgomez/sinai/>

## Motores de búsqueda

- Evaluación: cobertura (Search Engine Wars)
  - Equiparado a calidad (error!!!)



## Fuente SEW - 1000K documentos textuales indexados

Sistemas Inteligentes de Acceso a la Información – José María Gómez Hidalgo 38  
<http://www.esp.uem.es/~jmgomez/sinai/>

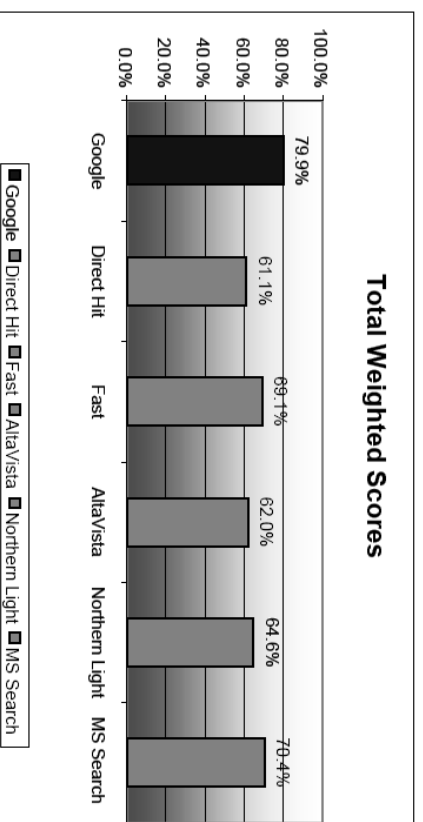
## Motores de búsqueda

- Evaluación: estudios de calidad
  - [eTestingLabs00]
    - Por encargo de Google, año 2000
    - Seis buscadores, 5 tipos de consultas X 5 temas, 10 primeros resultados
    - Puntuaciones según relevancia, enlaces vivos, bonus por acierto en primeros enlaces
    - No se considera la cobertura (imposible!!!)
    - Se mide porcentaje sobre puntos posibles

Sistemas Inteligentes de Acceso a la Información – José María Gómez Hidalgo  
<http://www.esp.uem.es/~jmgomez/sinai/> 39

## Motores de búsqueda

- Estudios de calidad
  - [eTestingLabs00] Resultados



Sistemas Inteligentes de Acceso a la Información – José María Gómez Hidalgo  
<http://www.esp.uem.es/~jmgomez/sinai/> 40

## Índice

- **Introducción**
- **Características de la Web**
- **Técnicas de búsqueda**
  - **Motores de búsqueda**
  - **Directorios**
  - Navegación
- Estudios de usuarios
- Resumen

Sistemas Inteligentes de Acceso a la Información – José María Gómez Hidalgo  
<http://www.esp.uem.es/~jmgomez/sinai/>

41

## Directorios

- Catálogos organizados de enlaces comentados
  - Mapas de carretera de la Web
  - Kioskos de enlaces = pegamento de la Web
- Similares a catálogos bibliográficos
  - Mantenidos por “expertos” en cada tema
  - Dudas en el proceso editorial (DMOZ vs Yahoo!)
- Punto de entrada **Web de alta calidad**

Sistemas Inteligentes de Acceso a la Información – José María Gómez Hidalgo  
<http://www.esp.uem.es/~jmgomez/sinai/>

42

## Directorios

- Inicio => Yahoo! (1995)
    - Iniciativa de dos jóvenes estudiantes
    - Algunas estadísticas (2000)
      - Más de 100 editores, 1.8M enlaces (sobre 2100M)
    - Reconvertido a portal generalista
      - Suite de servicios => correo, noticias, etc.
    - Hoy en día cotiza millones en bolsa
- <http://www.yahoo.com>, <http://dir.yahoo.com>

Sistemas Inteligentes de Acceso a la Información – José María Gómez Hidalgo  
<http://www.esp.uem.es/~jmgomez/sinai/>

43

## Directorios

- Un referente = Open Directory Project (ODP)
    - Esfuerzo colaborativo, licencia libre © Netscape
    - Algunas estadísticas
      - (2001) 36K editores, 361K categorías, 2.1M enlaces
      - (2006) 71K editores, 590K categorías, 5,2M enlaces
    - Directorio de referencia de Google
- <http://dmoz.org/>

Sistemas Inteligentes de Acceso a la Información – José María Gómez Hidalgo  
<http://www.esp.uem.es/~jmgomez/sinai/>

44

## Directorios

- Se puede (semi) automatizar la catalogación
  - Técnicas de Categorización Automática combinadas con Crawling
- Evaluación: objetivo precisión (no cobertura)
  - Demostrado que la contextualización de consultas y resultados ayuda en acceso a la información
- Derivan en portales corporativos
  - Definen el vocabulario corporativo, los puntos de interés, etc. a nuevos empleados + self-branding

Sistemas Inteligentes de Acceso a la Información – José María Gómez Hidalgo 45  
<http://www.esp.uem.es/~jimgomez/sinai/>

## Índice

- **Introducción**
- **Características de la Web**
- **Técnicas de búsqueda**
  - **Motores de búsqueda**
  - **Directorios**
  - **Navegación**
- Estudios de usuarios
- Resumen

Sistemas Inteligentes de Acceso a la Información – José María Gómez Hidalgo 46  
<http://www.esp.uem.es/~jimgomez/sinai/>

## Navegación

- Pulsar hiper-enlaces en la página actual
  - Búsqueda localizada, por objetivos, con evaluación previa del usuario
  - Combinada con las anteriores, deriva en un proceso complejo e impreciso
- Susceptible de diversas ayudas
  - Búsqueda localizada
  - Asistentes de navegación

Sistemas Inteligentes de Acceso a la Información – José María Gómez Hidalgo  
<http://www.esp.uem.es/~jimgomez/sinai/>

47

## Navegación

- Búsqueda localizada (WebGlimpse)
  - Búsqueda local de sitio Web, dependiente de la página y la vecindad
  - Proceso
    - Entrada = lista de páginas o sitio Web, páginas en las que se desea casilla de búsqueda, profundidad de la vecindad
    - Salida = (algunas) páginas con casillas de búsqueda cercana, vecindad local
    - Equivalente a seguir enlaces contruidos dinámicamente por búsqueda de vecinos <http://webglimpse.net/>

Sistemas Inteligentes de Acceso a la Información – José María Gómez Hidalgo  
<http://www.esp.uem.es/~jimgomez/sinai/>

48

## Navegación

- **Asistentes de navegación (Alexa Internet)**
  - Recomendadores de enlaces y sitios
    - Usa información de la página actual y adyacentes = enlaces, palabras clave
    - También usa información del usuario (histórico)
  - Actualmente proporcionada como barra
    - <http://www.alexa.com/>

Sistemas Inteligentes de Acceso a la Información – José María Gómez Hidalgo  
<http://www.esp.uem.es/~jmgomez/sinai/>

49

## Índice

- **Introducción**
- **Características de la Web**
- **Técnicas de búsqueda**
  - **Motores de búsqueda**
  - **Directorios**
  - **Navegación**
- **Estudios de usuarios**
- **Resumen**

Sistemas Inteligentes de Acceso a la Información – José María Gómez Hidalgo  
<http://www.esp.uem.es/~jmgomez/sinai/>

50

## Estudios de usuarios

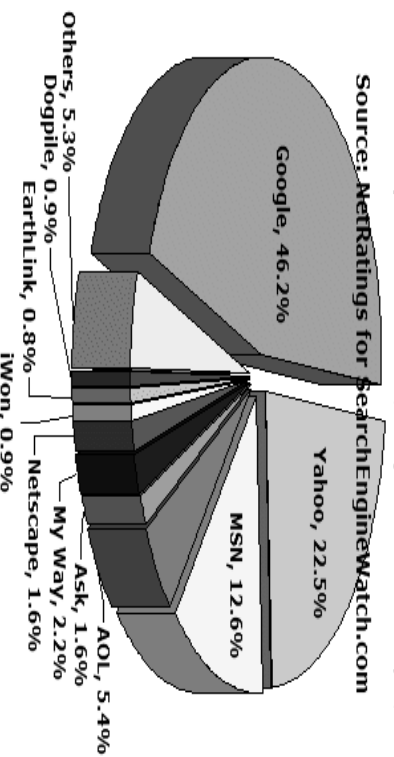
- Múltiples estudios de usuarios
  - De consultoras (Nielsen, etc.), disponibles en Search Engine Watch
    - Popularidad de buscadores
  - Científicos (Spink et al.) = Web log mining
    - Análisis estadístico detallado y fiable de logs de buscadores
    - Especialmente relevante [Spink02]

Sistemas Inteligentes de Acceso a la Información – José María Gómez Hidalgo  
<http://www.esp.uem.es/~jmgomez/sinai/>

51

## Estudios de usuarios

- Nielsen NetRatings para SEW (07/2005)
  - Monitorización de 1M usuarios con sistema propio
  - Resultados para USA (hogar + trabajo)



Sistemas Inteligentes de Acceso a la Información – José María Gómez Hidalgo  
<http://www.esp.uem.es/~jmgomez/sinai/>

52

# Estudios de usuarios

- [Spink02] Análisis de Web Logs de Excite
  - Metodología
    - 1M consultas, 200K usuarios, tres momentos (97, 99, 01)
    - Datos de los Web Logs

**Table 1. Excite data sets for 1997, 1999, and 2001.**

Data set	Sessions	Queries	Terms
1997 <sup>1</sup>	211,063	1,025,908	1,277,763
1999 <sup>2</sup>	325,711	1,025,910	1,500,500
2001	262,025	1,025,910	1,538,120

Sistemas Inteligentes de Acceso a la Información – José María Gómez Hidalgo 53  
<http://www.esp.uem.es/~jgomez/sinai/>

# Estudios de usuarios

- [Spink02] Análisis de Web Logs de Excite
  - Resultados

**Table 2. Comparative statistics for Excite Web query data sets— one million queries per study.**

Variables	1997	1999	2001
Mean terms per query	2.4	2.4	2.6
Terms per query			
1 term	26.3%	29.8%	26.9%
2 terms	31.5%	33.8%	30.5%
3+ terms	43.1%	36.4%	42.6%
Mean queries per user	2.5	1.9	2.3
Mean pages viewed per query	1.7	1.6	1.7
Pages viewed per query			
1 page	28.6%	42.7%	50.5%
2 pages	19.5%	21.2%	20.3%
3+ pages	51.9%	36.1%	29.2%
Users modifying queries	52.0%	39.6%	44.6%
Session size			
1 query	48.4%	20.8%	30.8%
2 queries	60.4%	19.8%	19.8%
3+ queries	55.4%	19.3%	25.3%
Boolean queries	5.0%	5.0%	10.0%
Terms not repeated in the data set	57.1%	61.6%	61.7%
Use of 100 most frequently occurring query terms	17.9%	19.3%	22.0%

Sistemas Inteligentes de Acceso a la Información – José María Gómez Hidalgo 54  
<http://www.esp.uem.es/~jgomez/sinai/>

# Estudios de usuarios

- [Spink02] Análisis de Web Logs de Excite
  - Resultados

Table 3. Distribution of query samples across general topic categories.

Rank	1997 Excite data set (2,414 queries)	1999 Excite data set (2,539 queries)	2001 Excite data set (2,453 queries)
1	19.9% Entertainment or recreation	24.5% Commerce, travel, employment, or economy	24.7% Commerce, travel, employment, or economy
2	16.8% Sex and pornography	20.3% People, places, or things	19.7% People, places, or things
3	13.3% Commerce, travel, employment, or economy	10.9% Computers or Internet	11.3% Non-English or unknown
4	12.5% Computers or Internet	7.8% Health or sciences	9.6% Computers or Internet
5	9.5% Health or sciences	7.5% Sex and pornography	8.5% Sex and pornography
6	6.7% People, places, or things	7.5% Entertainment or recreation	7.5% Health or sciences
7	5.7% Society, culture, ethnicity, or religion	6.8% Non-English or unknown	6.6% Entertainment or recreation
8	5.6% Education or humanities	5.3% Education or humanities	4.5% Education or humanities
9	5.4% Performing or fine arts	4.2% Society, culture, ethnicity, or religion	3.9% Society, culture, ethnicity, or religion
10	4.1% Non-English or unknown	1.6% Government	2.0% Government
11	3.4% Government	1.1% Performing or fine arts	1.1% Performing or fine arts

Sistemas Inteligentes de Acceso a la Información – José María Gómez Hidalgo 55  
<http://www.esp.uem.es/~jmgomez/sinai/>

## Índice

- **Introducción**
- **Características de la Web**
- **Técnicas de búsqueda**
  - **Motores de búsqueda**
  - **Directorios**
  - **Navegación**
- **Estudios de usuarios**
- **Resumen**

Sistemas Inteligentes de Acceso a la Información – José María Gómez Hidalgo 56  
<http://www.esp.uem.es/~jmgomez/sinai/>

## Resumen

- La Web se ha convertido en un recurso de información extremadamente popular
- Por múltiples razones, es el entorno más retador para el acceso a la información
- Hasta el momento, sólo las técnicas más simples sobreviven
  - Extensión de las técnicas habituales