Matemáticas de la Web:

Google, robots y monstruos

F. Javier Soria de Diego

Universidad de Barcelona

- ¿Qué es la World Wide Web (WWW, la Web o Red Global Mundial)?
- ¿Qué forma, tamaño o estructura tiene la Web?
- ¿Qué es Google?
- ¿Qué propiedades matemáticas se utilizan cuando «navegamos»?

La Web tiene sus orígenes en los años 1990 (la propuesta formal se publicó el 12 de noviembre de 1990 y el primer servidor fue puesto en línea el 6 de agosto de 1991) y se debe en gran medida a la labor de Timothy Berners-Lee (Tim) del CERN.

Tim Berners-Lee,
el «padre» de la Web

Tim es licenciado en física, hijo de matemáticos y posee todo tipo de distinciones en reconocimiento a su trabajo (es Sir, oficial de la Orden del Imperio Británico, Premio Príncipe de Asturias de Investigación Científica y Técnica en 2002, Doctor Honoris Causa por diversas Universidades, miembro de Sociedades y Academias Científicas).

Primer servidor Web: NeXTcube, 1991

Servidor de la Wikipedia

En 1992 había 26 servidores Web. En 2009, 110 millones.

Uno de los primeros navegadores desarrollados en la Web fue Mosaic:

Actualmente existen muchos otros:

Una primera definición muy básica nos diría que la Web hace referencia a las páginas a las que nos conectamos con nuestro navegador (la información, el software), mientras que Internet es la red de ordenadores y servidores donde se almacenan (la parte hardware).

Una página de la Web está escrita en un lenguaje llamado HTML (HyperText Markup Language) o lenguaje de etiquetas de hipertexto.

Estás páginas se almacenan en una red de ordenadores que se comunican entre sí siguiendo una serie de instrucciones o protocolo: HTTP (HyperText Transfer Protocol).

Toda página tiene su propia dirección que la identifica de manera única: URL (Uniform Resource Locator).

Código HTML

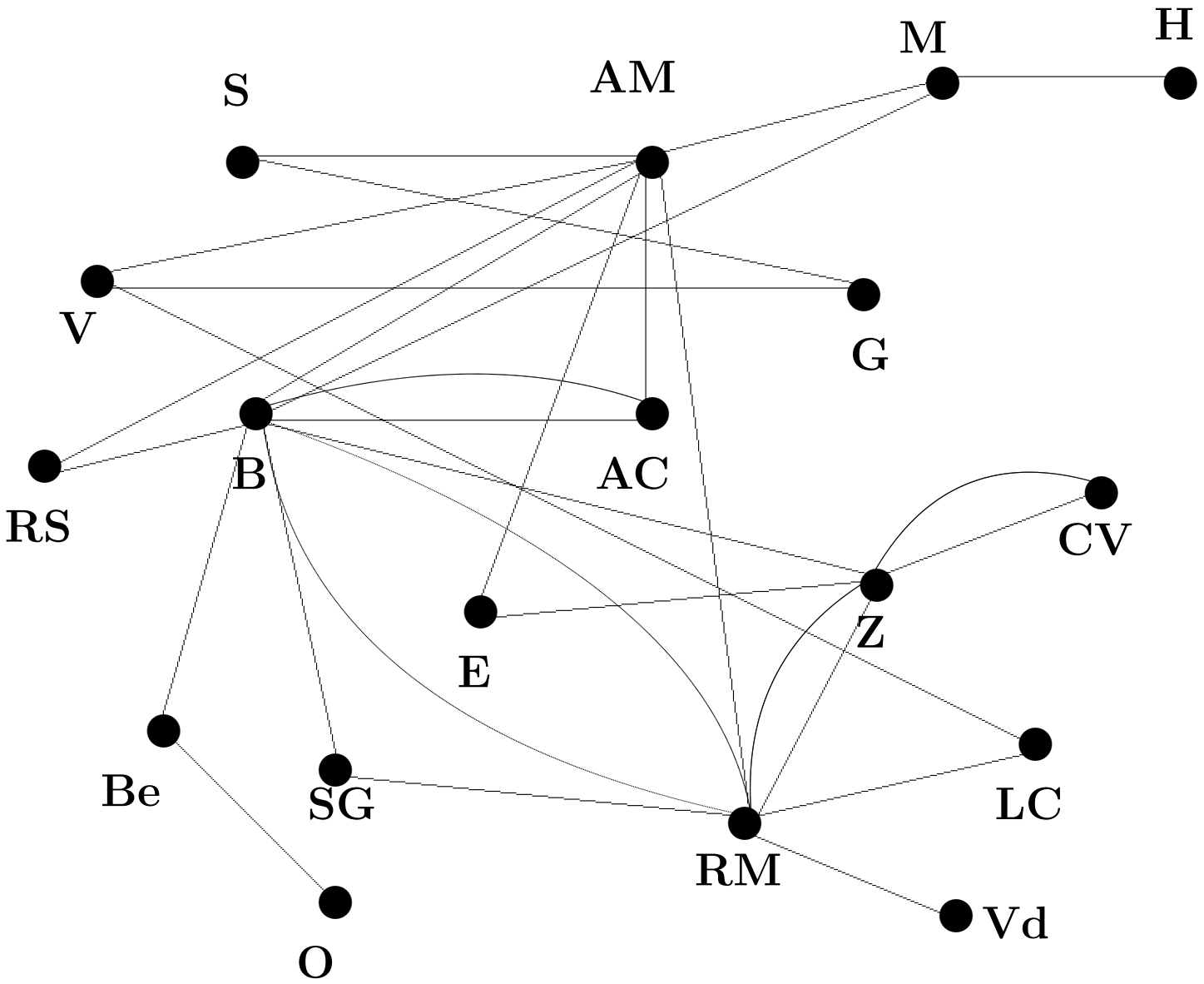
```
<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.0 Transitional//EN">
<html>
<head>
<STYLE type="text/css">
.text {
font-family: Verdana, Arial, Helvetica, sans-serif;
font-size: 12px;}
A.menu {
color: #FFFFFF;}
</STYLE>
<title>Prueba</title>
</head>
<body>
Esto es una prueba: vfrtgbnhy12345
</body>
</html>
```

El navegador nos muestra la página Web titulada «Prueba» cuya URL es http://garf.ub.es/prueba.html

La manera habitual de modelizar una red, como puede ser la Web, es mediante el uso de grafos, que no son más que una colección de puntos (o vértices) unidos entre sí, siguiendo alguna regla establecida, mediante líneas (o aristas).

Grafo de la red de estaciones del metro de Barcelona:

Finalistas de la Copa del Rey de fútbol: 1980-2010



¿Cómo se representa la Web?

Siguiendo el modelo de un grafo, los vértices representarán todas las páginas (o direcciones URL), y dos páginas (o vértices) estarán unidas, si tienen un enlace (link) en común:

Pero... ¿cuántas páginas existen en la Web?

La respuesta nunca puede ser exacta, pues la Web está en continua evolución (aparecen y desaparecen las páginas constantemente).

Un estudio de 2001 muestra que había del orden de

550.000.000.000

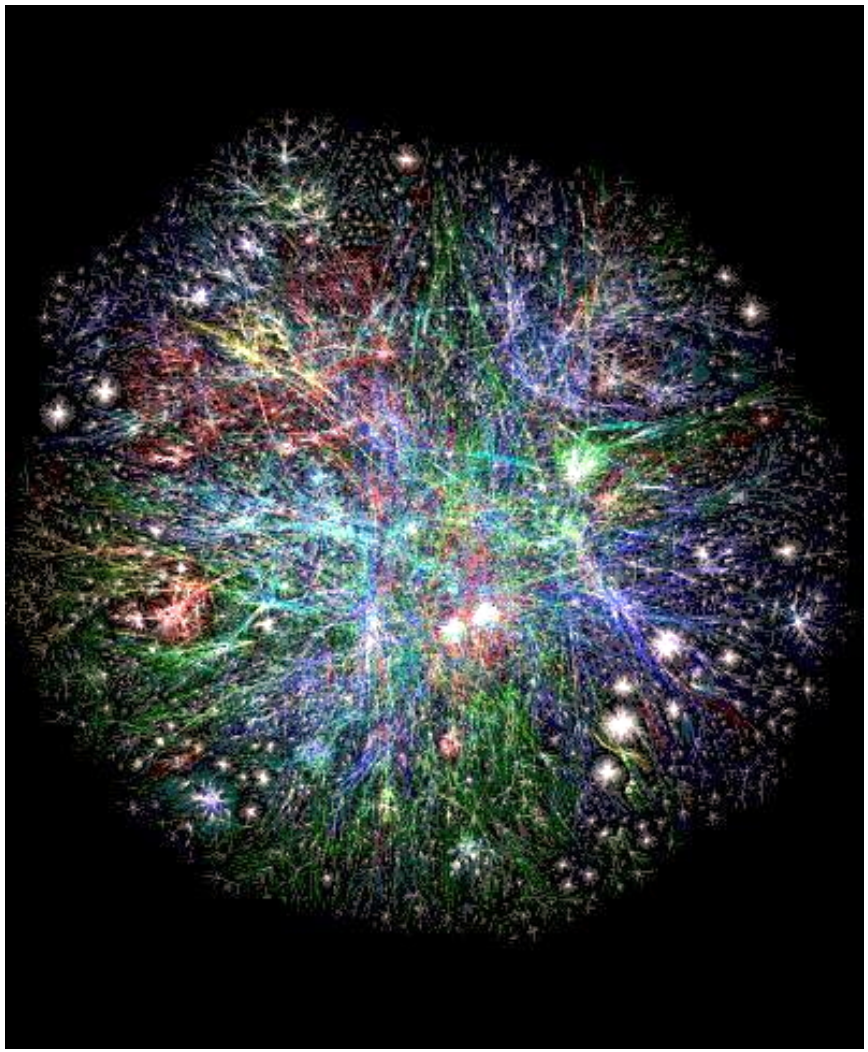
(550 mil millones) de páginas en la Web.

En 2008 este número se incrementó a

1.000.000.000.000

(1 billón).

Así, no es de extrañar que el grafo de la Web sea extremadamente complicado de dibujar...



Representación gráfica de la WWW.

Este dato nos plantea un par de preguntas interesantes:

1. **¿Es este grafo un «monstruo» aleatorio, o existen algunas propiedades que se puedan describir desde un punto de vista matemático, y que nos permitan hacer un análisis de su estructura y evolución?**
2. **¿Cómo podemos encontrar la información que nos interesa entre tal cantidad de páginas?**

   Un pequeño cálculo: si tuviésemos que revisar una a una todas las páginas (1 billón), y tardáramos un segundo por página, nos llevaría unos 31.700 años.

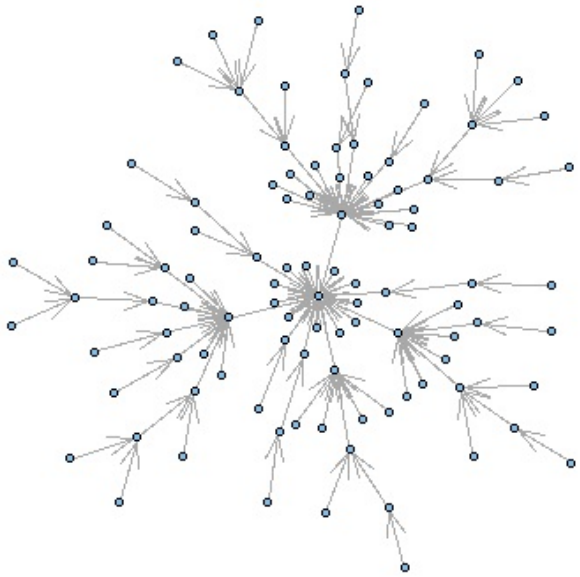
. . . . . .

Google

El modelo que explica las principales propiedades de la Web se debe a los matemáticos A. Barabási y R. Albert (1998), y está basado en la llamada ley de unión preferencial:

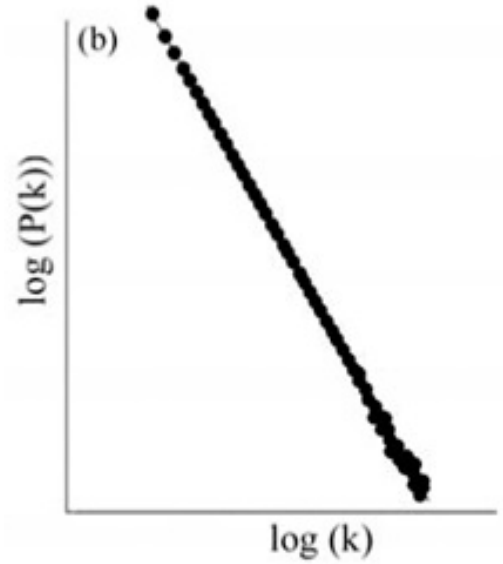
Las nuevas páginas que aparecen en la Web, se enlazan a otras páginas eligiendo aquéllas que ya tienen muchos enlaces (esto se conoce como que «los ricos se hacen más ricos».

Por ejemplo, es más fácil que aparezca un enlace a la Web del Barça que a la del equipo de mi pueblo.



Algunas propiedades de esta construcción son:

- **El número de aristas de cada vértice sigue una distribución de «ley de potencia» k^{α}, $-3 < \alpha < -1$:**



  Muchos vértices tienen pocas aristas, y unos pocos muchas.

- **El diámetro del grafo es muy pequeño en comparación al número de vértices: «mundo pequeño» (se calcula que en la población humana, dos personas cualesquiera acaban relacionándose a través de una cadena de 6 personas que se conocen de manera sucesiva).**

Datos de la Web:

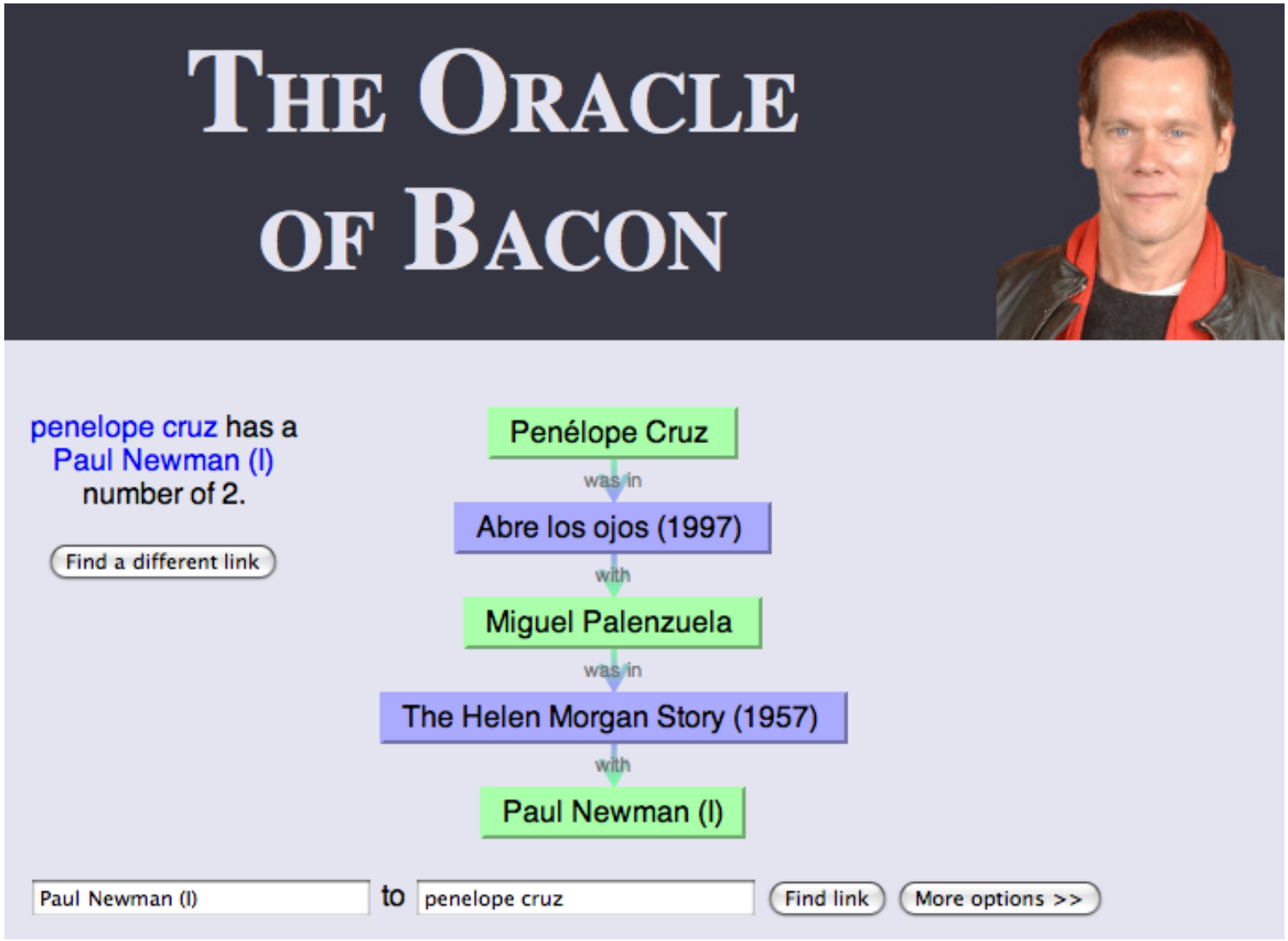
- **Sigue un modelo de ley de potencia: $k^{-2,45}$.**
- **Mundo pequeño:**

  Se calcula que el número de páginas que hay que visitar para ir de una página cualquiera a otra es del orden de 19 clicks.

Estas propiedades se han observado en grafos que modelizan todo tipo de redes: sociales, ecosistemas de poblaciones, proteínas, etc.

Veamos un ejemplo concreto:

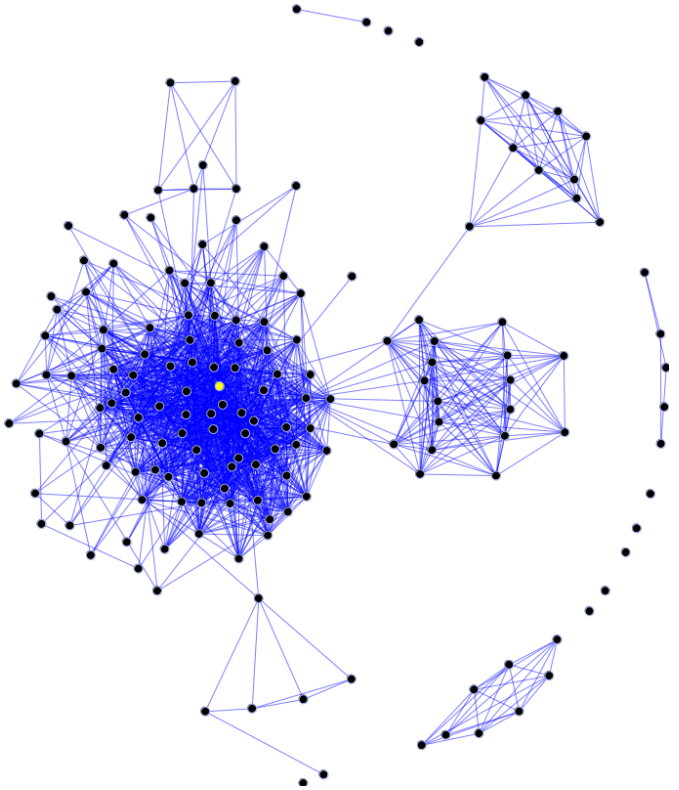
Grafo de Hollywood



http://oracleofbacon.org/cgi-bin/movielinks

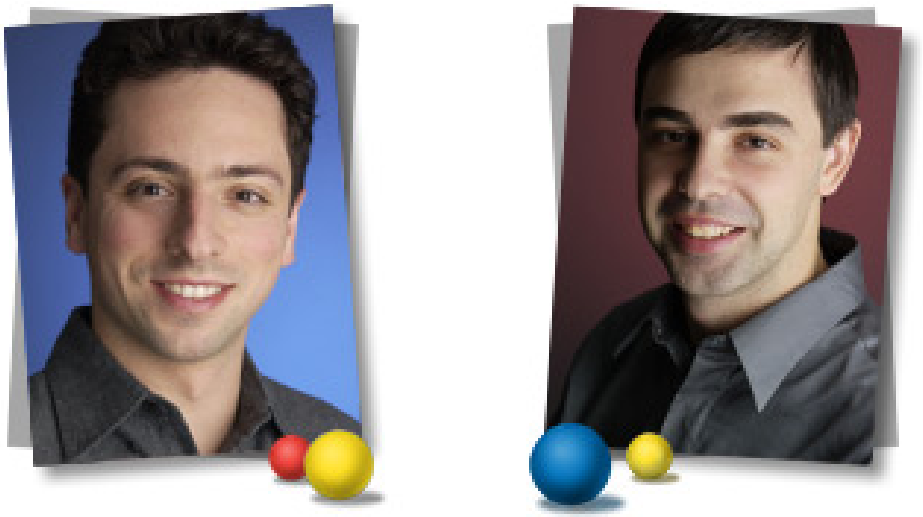
Kevin Bacon Number	# of People
0	1
1	2108
2	204188
3	601747
4	136178
5	8656
6	839
7	111
8	12

Número de actores



El grafo de Kevin Bacon

El diámetros de este grafo es 12 y el exponente de la ley de la potencia es -2,3.



Sergey Brin y Larry Page

Google fue fundada por Sergey Brin y Larry Page en 1998, y tiene como objetivo «... organizar la información mundial para que resulte universalmente accesible y útil».

Sergey y Larry nacieron ambos en 1973, y son titulados en Matemáticas y Ciencias de la Computación.

Google, que se lee «Gúgol», es un término matemático que se utiliza para referirse a un 1 seguido de 100 ceros = 10^{100}**.**

Google se ha convertido en la empresa líder en Internet, gracias principalmente a su motor de búsqueda que ha supuesto una auténtica revolución en la sociedad del siglo XXI.

Cuenta con cerca de 20.000 empleados repartidos por todo el mundo, y es una de las empresas más valiosas del mercado financiero (en 2009 obtuvo un beneficio de 4.600 millones de euros).

Fue galardonada con el Premio Príncipe de Asturias de Comunicación y Humanidades de 2008.

Google ofrece en la actualidad toda clase de servicios:

Google Search, Gmail, Google Maps, Google News, Google SketchUp, Google Video, Google Earth, AdSense, AdWords, YouTube, Picasa, Google Chrome, Google Analytics, Google Buzz, Google Desktop, ...

Buscar

Académico
Busca documentos académicos

Alertas
Recibe noticias y resultados de búsquedas por correo electrónico

Barra Google
Añade un cuadro de búsqueda a tu navegador

Búsqueda de blogs
Busca blogs sobre tus temas favoritos

Búsqueda en la web
Busca millones de paginas web

Búsqueda personalizada
Crea una experiencia de búsqueda personalizada para tu comunidad

Google Chrome
Un navegador que ofrece rapidez, estabilidad y seguridad

Desktop
Realiza búsquedas en tu propio equipo

Earth
Una ventana tridimensional al planeta

Funcionalidades de búsqueda web
Saca el máximo partido a tus búsquedas

iGoogle
Añade noticias, juegos y mucho más a tu página de inicio personalizada de Google

Imágenes
Busca imágenes en la web

Libros
Realiza búsquedas en libros

Maps
Direcciones y directorio de negocios

Noticias
Busca miles de noticias

Videos
Busca vídeos y carga los tuyos

Explorar e innovar

Code
Herramientas, APIs y recursos para programadores

Labs
Prueba nuevos productos Google

Comunicar, mostrar y compartir

Blogger
Expresa tus opiniones online

Calendar
Organiza tu agenda y comparte eventos con tus amigos

Docs
Crea tus proyectos online, compártelos y accede a ellos desde donde estés.Crea tus proyectos online, compártelos y accede a ellos desde donde estés

Gmail
Correo rápido, con menos spam y con la tecnología de búsqueda de Google

Grupos
Crea listas de distribución y grupos de debate

Picasa
Encuentra, edita y comparte tus fotografías

Google Reader
Recibe los feeds de tus blogs y sitios de noticias favoritos

Sites
Crea sitios web y wikis de grupos seguras

SketchUp
Construye modelos 3D de forma rápida y fácil

Talk
Envía mensajes instantáneos y llama a tus amigos desde tu equipo

Traductor
Visualiza páginas web en otros idiomas

YouTube
Ve, sube y comparte vídeos

Apuntarse a la movilidad

Mobile
Usa Google desde tu teléfono móvil

Maps para móviles
Consulta mapas y obtén itinerarios en tu teléfono móvil

Búsqueda
Busca en Google desde tu teléfono móvil

Optimizar el funcionamiento del equipo informático

Pack
Una colección gratuita de software indispensable

Pero, ¿cómo consigue Google toda la información?, ¿cómo la ordena?, ¿y cómo selecciona los «mejores» resultados?

Google dispone de un conjunto de programas informáticos, llamados «Googlebots», que rastrean (buscan) continuamente miles de millones de páginas en la Web. Estos programas son los llamados «Robots», que de manera silenciosa «espían» con una velocidad de ejecución vertiginosa toda la Web.

El proceso de rastreo que Google lleva a cabo empieza con una lista de diferentes URL. El Googlebot hace una copia de cada una de las páginas que rastrea con el fin de compilar un índice masivo de todas las palabras que reconoce (se calcula que hay unas 25 mil millones de páginas indexadas, un 2,5 % del total).

Cuando un usuario realiza una consulta, Google busca en el índice las páginas que coinciden con la búsqueda y devuelve al usuario los resultados más relevantes (suele haber miles de resultados, y casi siempre nos da primero el que queremos...).

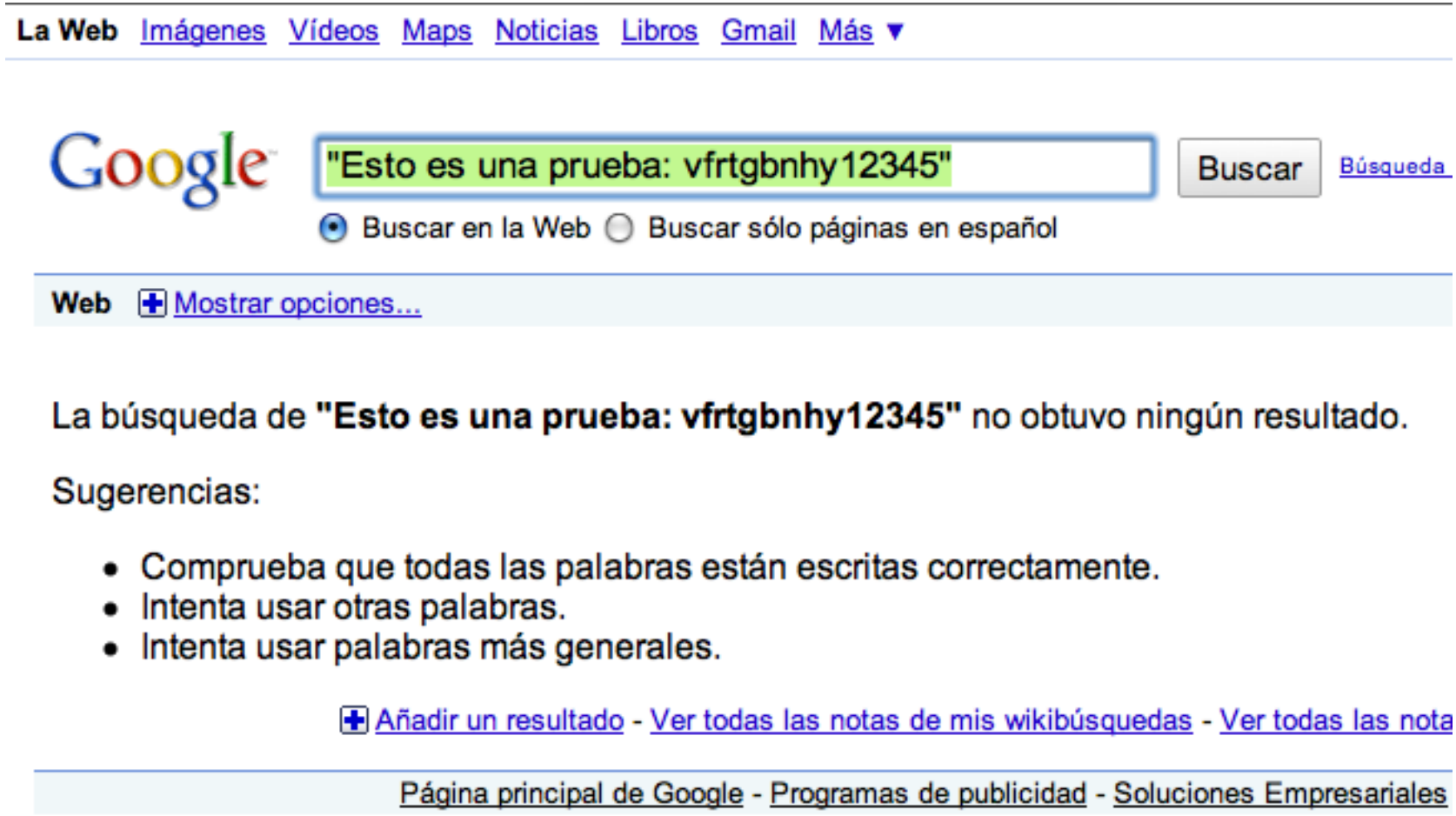
La relevancia se determina en función de unos 200 factores, uno de los cuales es el «PageRank» de cada página. El PageRank es la medida de la importancia de una página en función de las veces que la citan otras páginas.

Recogida de datos y creación del índice . . . . . Googlebots

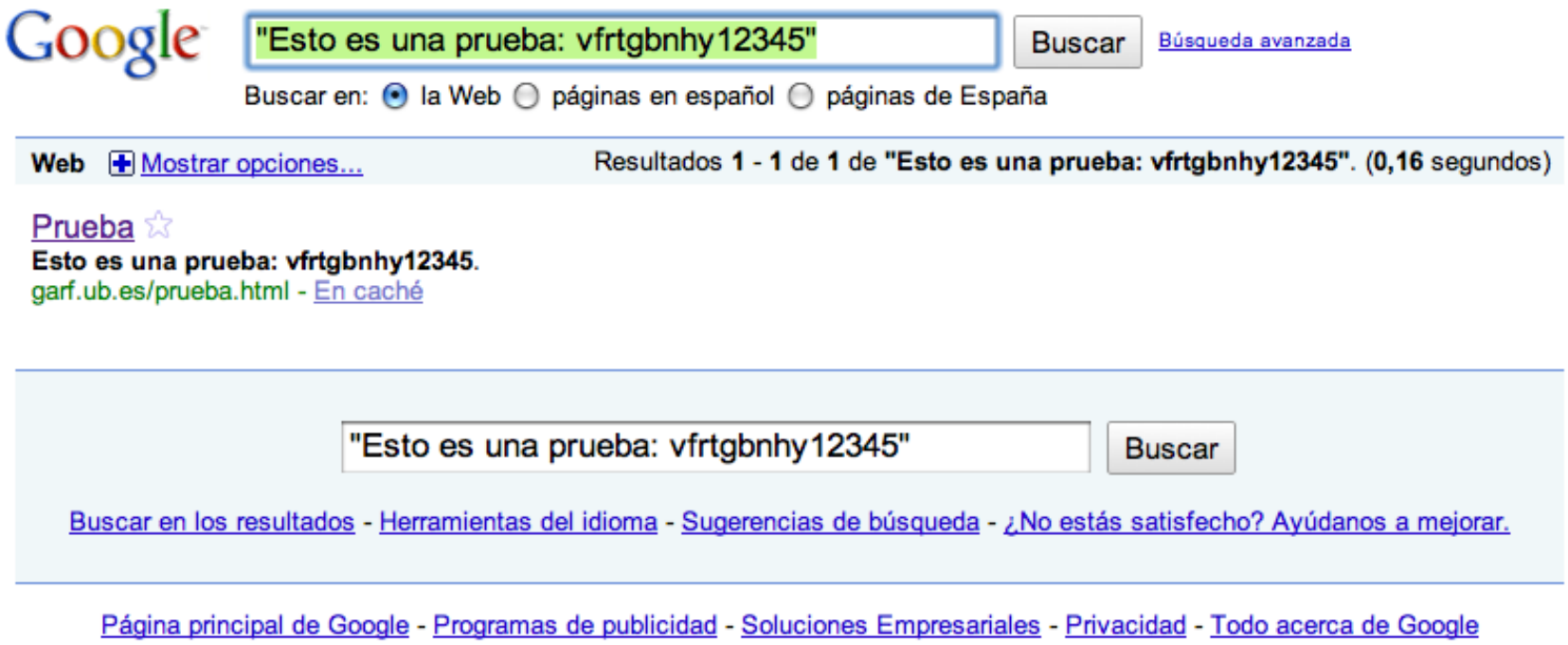
Selección de los mejores resultados . . . . . PageRank

Distintas versiones de los Googlebots...

¿Cuánto tarda Google en indexar la Web? Veamos un ejemplo:



El día 8 de febrero empezamos la búsqueda de nuestra página de prueba...



... hasta que el 24 de febrero (16 días después) Google la encuentra.

Una estimación muy preliminar es que los robots de Googlebots revisan unas ¡18.000 páginas por segundo!

Muy bien, ahora ya tenemos almacenadas todas las páginas de la Web. ¿Qué hacemos con ellas? ¿Cómo las ordenamos en orden de prioridad? ¿Cuáles son las mejores?

De entre todas las que hablan de la tortilla de patatas (682.000), ¿cómo le damos al usuario aquéllas que le van a resultar de mayor interés?

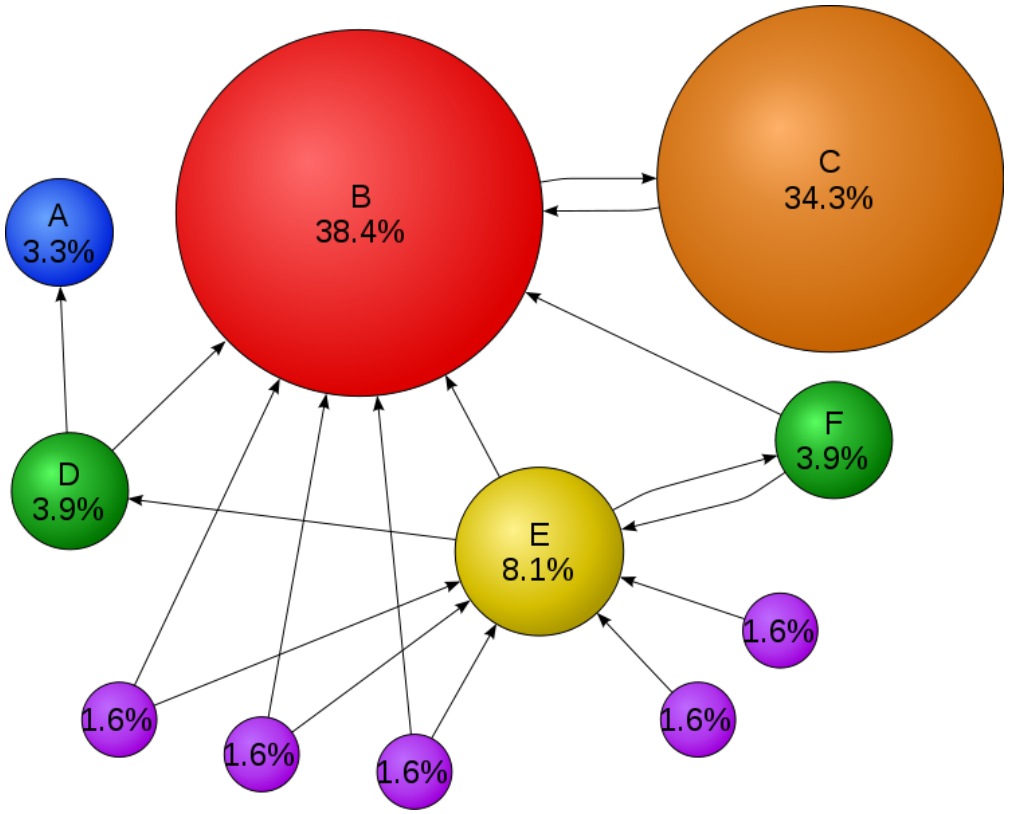
Esto es lo que se consigue mediante el número PageRank, y que vamos a describir brevemente. Hay que reseñar que la formulación y desarrollo de este método es la piedra angular de todo el desarrollo de Internet de la última década, y puede decirse que la Web existe gracias a la invención de PageRank:

«The Anatomy of a Large-Scale Hypertextual Web Search Engine», Sergey Brin y Lawrence Page (1998).

http://infolab.stanford.edu/~backrub/google.html

El valor de PageRank de una página se calcula teniendo en cuenta su popularidad, y ésta se mide en función de cuántas veces es citada por otras páginas (los votos que recibe).

Pero los votos no valen lo mismo: son mejores si provienen de páginas que tienen un PageRank alto.



Da la impresión de que para calcular el PageRank de una página, hay que conocer primero el PageRank de las páginas que la citan... y esto tiene pinta de que es un bucle que nos lleva al punto de partida.

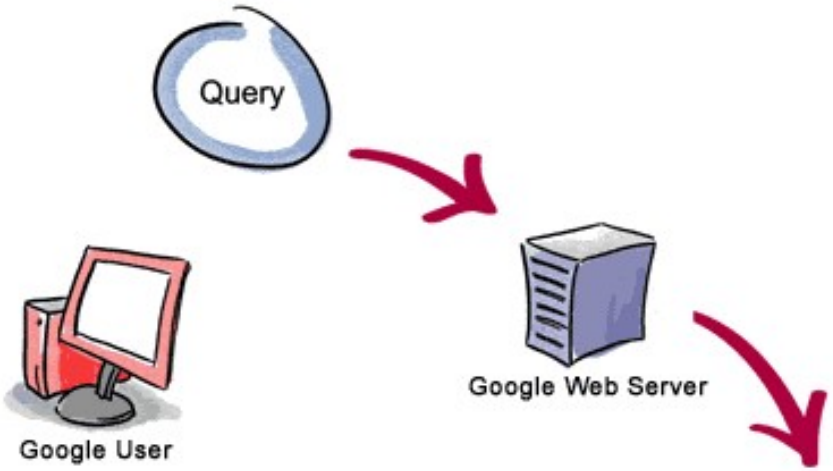
Y esto es verdad. Pero afortunadamente, la solución a este problema nos lo dan las Matemáticas, y equivale a resolver un sistema de ecuaciones lineales:

Si $PR(A)$ es el PageRank de la página A y $B, C, D, ...$ son las N páginas que citan a A (es decir, que tienen un enlace a dicha página), y si $N(B), N(C), N(D), ...$ es el número de enlaces que aparecen respectivamente en esas páginas, entonces:

$$PR(A) = \frac{1-d}{N} + d\left(\frac{PR(B)}{N(B)} + \frac{PR(C)}{N(C)} + \frac{PR(D)}{N(D)} + \cdots\right)$$

donde d es un parámetro que se elige alrededor del valor 0,85.

Aunque $0 \leq PR(A) \leq 1$, normalmente se escala para que sea un número del 0 al 10.



1. El servidor web envía la consulta a los servidores de indexación. El contenido que se encuentra en los servidores de indexación es similar al índice de un libro: indica qué páginas contienen las palabras que coinciden con la consulta.

Googlebots

2. La consulta viaja hasta los servidores de documentos, en los cuales se recuperan los documentos almacenados. Se generan fragmentos para describir cada resultado de búsqueda.

PageRank

3. El usuario recibe los resultados de su búsqueda en una fracción de segundo.

El PageRank de los 3 primeros resultados es de 4, el del puesto número 200 es de 2, el del puesto número 700 es de 1...

1.- Barcelona: 10

2.- Atlético de Madrid: 8

2.- Real Madrid: 8

4.- Zaragoza: 6

5.- Athletic de Bilbao: 3

5.- Mallorca: 3

5.- Valencia: 3

8.- Betis: 2

8.- Celta de Vigo: 2

8.- Deportivo de La Coruña: 2

8.- Espanyol: 2

8.- Getafe: 2

8.- Real Sociedad: 2

8.- Sevilla: 2

8.- Sporting de Gijón: 2

16.- Huelva: 1

16.- Osasuna: 1

16.- Valladolid: 1

Clasificación Copa del Rey según PageRank

0.148612- Barcelona: 10

0.122601- At. de Madrid: 8

0.119571- Real Madrid: 8

0.089469- Zaragoza: 6

0.054105- Mallorca: 3

0.053686- Valencia: 3

0.046623- Ath. Bilbao: 3

0.043911- Betis: 2

0.039813- Getafe: 2

0.038280- Sevilla: 2

0.036248- Depor: 2

0.034034- Espanyol: 2

0.033991- R. Sociedad: 2

0.033683- Celta de Vigo: 2

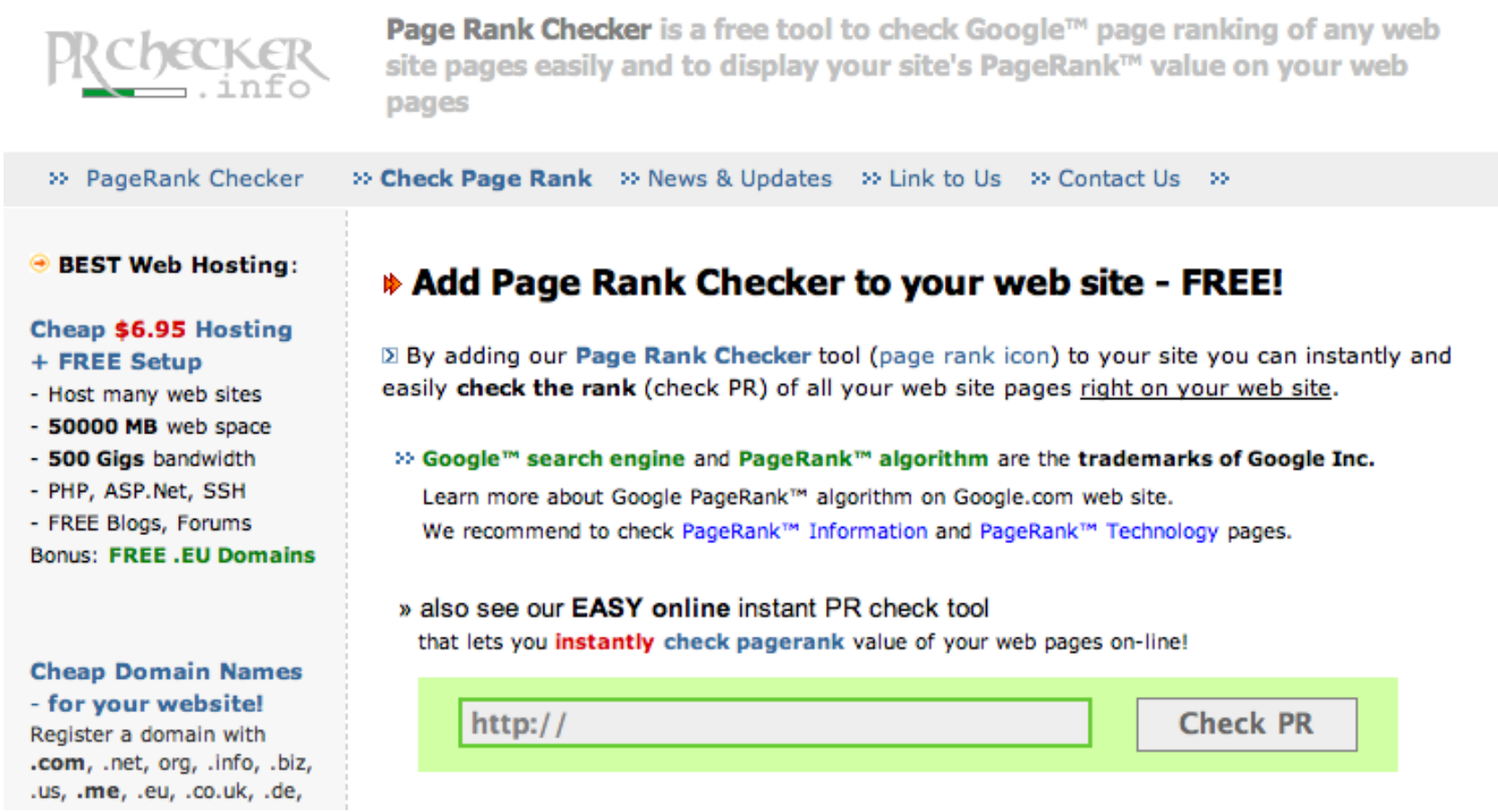
0.033669- Sporting: 2

0.026995- Osasuna: 1

0.023663- Huelva: 1

0.021037- Valladolid: 1

Cálculo del PageRank de una página web



http://www.prchecker.info/

Direcciones	PageRank
Facebook	10
Google	10
Yahoo	9
YouTube	9
Wikipedia	9
MSN	8

UAB	8
UB	8
UPC	8
UPF	8

Direcciones	PageRank
Arsenal	7
Espanyol	7
FC Barcelona	7
Manchester United	7
R. Madrid	7
Inter Milán	6
Bayern Munich	5
Nàstic	5
Alcorcón	3
Sant Cugat	3

RESUMEN

En la UB, por la mañana
he estado en la MateFest,
me han hablado de una telaraña
y de todo lo que hay en Internet.

Entre http y html,
url, Google y PageRank,
he visto, como si fuera la tele,
fútbol, actores y el logaritmo de k.

[Anónimo...]