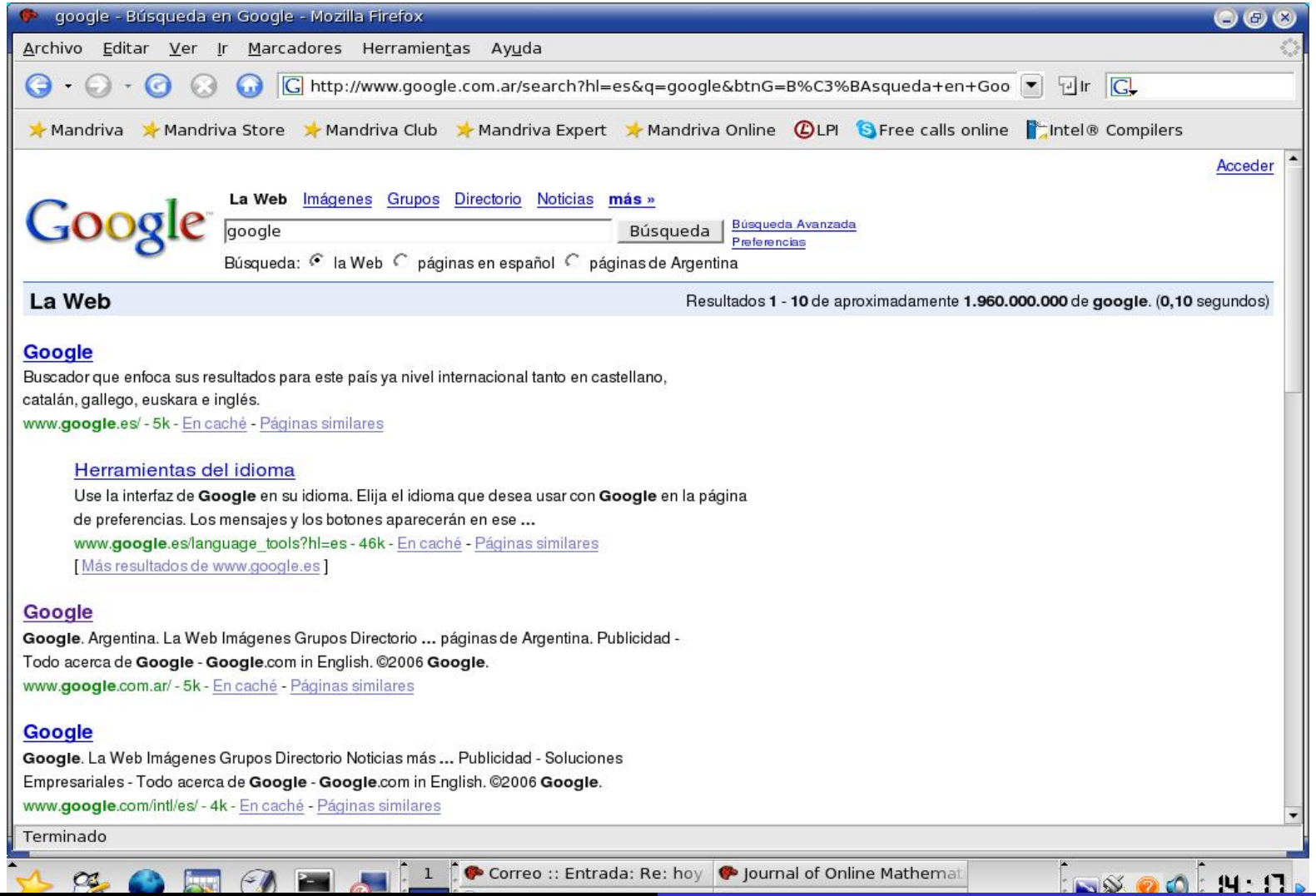
Buscando "Google" en Google

Gabriel Acosta

April 27, 2007

WWW.GOOGLE.COM



Algunos detalles

- Numero de paginas! $= 1.900.000.000$.

Algunos detalles

- Numero de paginas! = 1.900.000.000.
- Tiempo de la busqueda = 0.10 segundos!.

Algunos detalles

- Numero de paginas! $=$ 1.900.000.000.
- Tiempo de la busqueda $=$ 0.10 segundos!.
- El paginado es cada 10 paginas: visitando 1 por segundo ... aproximadamente $6,3$ años.

Algunos detalles

- Numero de paginas! $=$ 1.900.000.000.
- Tiempo de la busqueda $=$ 0.10 segundos!.
- El paginado es cada 10 paginas: visitando 1 por segundo ... aproximadamente $6,3$ años.
- Parece que solo muestran los 1000 primeros (quien tiene tanta paciencia!)

Algunos detalles

- Numero de paginas! $= 1.900.000.000$.
- Tiempo de la busqueda $= 0.10$ segundos!.
- El paginado es cada 10 paginas: visitando 1 por segundo ... aproximadamente $6,3$ años.
- Parece que solo muestran los 1000 primeros (quien tiene tanta paciencia!)
- Cuantos "Juanes" hay????

Algunos detalles

- Numero de paginas! $= 1.900.000.000$.
- Tiempo de la busqueda $= 0.10$ segundos!.
- El paginado es cada 10 paginas: visitando 1 por segundo ... aproximadamente $6,3$ años.
- Parece que solo muestran los 1000 primeros (quien tiene tanta paciencia!)
- Cuantos "Juanes" hay????
- Criterios de presentación de la información.

La Importancia de las Cosas ...

1. Importancia de los contenidos.

La Importancia de las Cosas ...

1. Importancia de los contenidos.
2. Importancia de las páginas

La Importancia de Los Contenidos

- Pagina de la Biblioteca Digital de la Mathematical Association of America
  http://mathdl.maa.org/

La Importancia de Los Contenidos

- Pagina de la Biblioteca Digital de la Mathematical Association of America
  http://mathdl.maa.org/
- Supongamos que tenemos siete páginas $P_1, P_2, ..., P_7$ con contenidos sobre "Postres", "Panes" y "Vegetales".

La Importancia de Los Contenidos

- Pagina de la Biblioteca Digital de la Mathematical Association of America
  http://mathdl.maa.org/
- Supongamos que tenemos siete páginas $P_1, P_2, ..., P_7$ con contenidos sobre "Postres", "Panes" y "Vegetales".
- Aunque no todas tienen la misma cantidad de información sobre cada item. Por ejemplo P_1 se dedica solo a "Postres", mientras que por ejemplo P_3 dedica el 70% a "Panes" y el 30% a "Vegetales" sin dar información sobre "Postres"

La Importancia de Los Contenidos

- Pagina de la Biblioteca Digital de la Mathematical Association of America
  http://mathdl.maa.org/
- Supongamos que tenemos siete páginas $P_1, P_2, ..., P_7$ con contenidos sobre "Postres", "Panes" y "Vegetales".
- Aunque no todas tienen la misma cantidad de información sobre cada item. Por ejemplo P_1 se dedica solo a "Postres", mientras que por ejemplo P_3 dedica el 70% a "Panes" y el 30% a "Vegetales" sin dar información sobre "Postres"
- Cómo podemos cuantificar esto?.

Vectorizamos los datos

- Podemos tomar para P_1 y P_3

$$\begin{matrix} \text{Postres} \\ \text{Panes} \\ \text{Vegetales} \end{matrix} \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} \qquad \begin{matrix} \text{Postres} \\ \text{Panes} \\ \text{Vegetales} \end{matrix} \begin{pmatrix} 0 \\ 0.7 \\ 0.3 \end{pmatrix}$$

Vectorizamos los datos

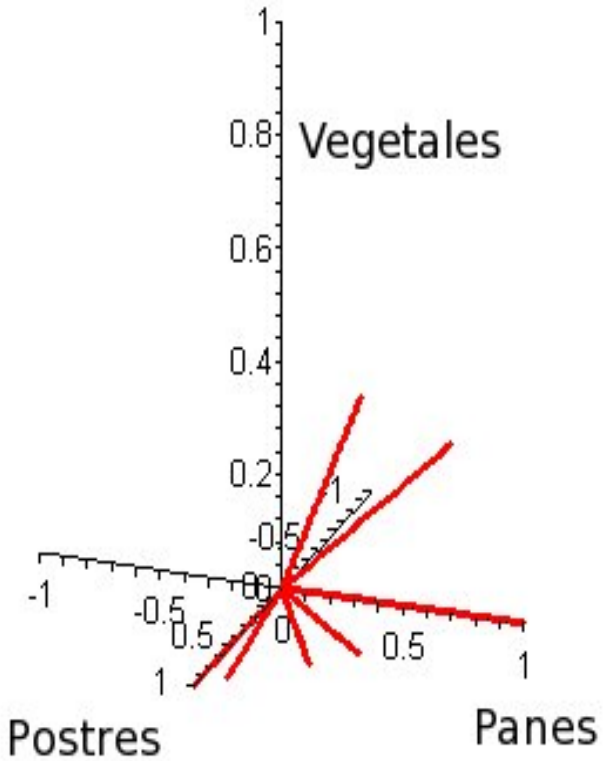
- Podemos tomar para P_1 y P_3

$$\begin{array}{c} \text{Postres} \\ \text{Panes} \\ \text{Vegetales} \end{array} \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} \qquad \begin{array}{c} \text{Postres} \\ \text{Panes} \\ \text{Vegetales} \end{array} \begin{pmatrix} 0 \\ 0.7 \\ 0.3 \end{pmatrix}$$

- Si hacemos lo propio con las restantes páginas podemos juntar la información en una "matriz"

$$\begin{array}{c} \text{Tipo de Comida} \\ \text{Postres} \\ \text{Panes} \\ \text{Vegetales} \end{array} \begin{array}{c} P_1 \quad P_2 \quad \ldots \quad P_7 \\ \begin{pmatrix} 1 & 0 & \ldots & 0.9 \\ 0 & 1 & \ldots & 0.1 \\ 0 & 0 & \ldots & 0 \end{pmatrix} \end{array}$$

Interpretación Geometrica

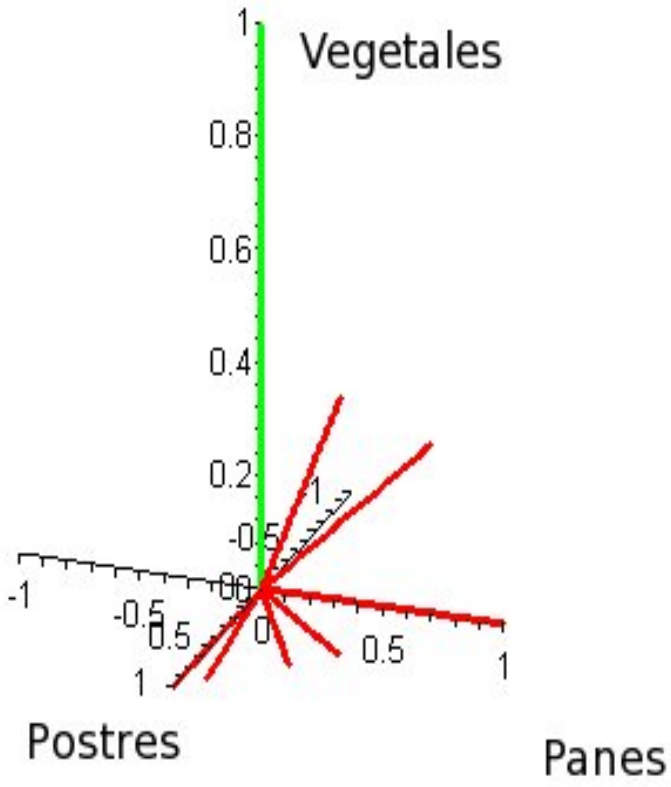


Busqueda "Geometrica"

Imaginemos que buscamos información sobre "Vegetales". Nuestro criterio vectorizado será

$$\begin{matrix} \text{Postres} \\ \text{Panes} \\ \text{Vegetales} \end{matrix} \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$$

Graficamente



Busqueda con un vector general

- Si tenemos un vector general V con el criterio que nos interesa

Busqueda con un vector general

- Si tenemos un vector general V con el criterio que nos interesa
- $$\begin{matrix} \text{Postres} \\ \text{Panes} \\ \text{Vegetales} \end{matrix} \begin{pmatrix} a \\ b \\ c \end{pmatrix}$$

Busqueda con un vector general

- Si tenemos un vector general V con el criterio que nos interesa
- $$\begin{matrix} \text{Postres} \\ \text{Panes} \\ \text{Vegetales} \end{matrix} \begin{pmatrix} a \\ b \\ c \end{pmatrix}$$
- Como elegimos las páginas relevantes?

Cuestión de ángulos

- Fijamos un cierto ángulo α

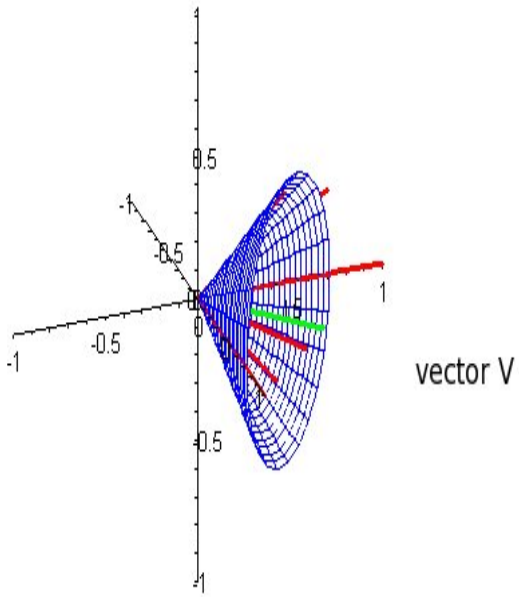
Cuestión de ángulos

- Fijamos un cierto ángulo α
- Y elegimos las páginas cuyo vector forme un ángulo menor a α respecto de V.

Cuestión de ángulos

- Fijamos un cierto ángulo α
- Y elegimos las páginas cuyo vector forme un ángulo menor a α respecto de V.
- Dicho de otro modo: construimos un "cono" con eje en V y elegimos las paginas que caen dentro del cono

Graficamente



Orden????

Damos las páginas (vectores) dentro del cono ordenados por ángulos crecientes!.

En la realidad ...

- Elección cuidadosa del α

En la realidad ...

- Elección cuidadosa del α
- Un α chico nos deja con una busqueda muy estricta (tal vez sin resultados!), un α grande con demasiados resultados!!!

En la realidad ...

- Elección cuidadosa del α
- Un α chico nos deja con una busqueda muy estricta (tal vez sin resultados!), un α grande con demasiados resultados!!!
- Consideremos **todas** las palabras del diccionario (en inglés hay cerca de 300.000).

En la realidad ...

- Elección cuidadosa del α
- Un α chico nos deja con una busqueda muy estricta (tal vez sin resultados!), un α grande con demasiados resultados!!!
- Consideremos **todas** las palabras del diccionario (en inglés hay cerca de 300.000).
- Ordenadas alfabéticamente de "aarónico" a "zuzón". Una busqueda por las palabras "deportes" y "playa" puede vectorizarse como hicimos antes ...

Vector de busqueda

$$\begin{array}{c} \text{aarónico} \\ \vdots \\ \text{deportes} \\ \vdots \\ \text{playa} \\ \vdots \\ \text{zuzón} \end{array} \begin{pmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{pmatrix}$$

Muchas cuentas para hacer!!!

Paginas en Internet $\sim$ 3.000.000.000

$$G = \left\{ \begin{array}{c} \text{Palabras} \\ \text{palabra 1} \\ . \\ . \\ . \\ . \\ . \\ . \\ . \\ . \\ \text{palabra 300.000} \end{array} \right. \begin{array}{c} \begin{array}{cccc} \text{Pag1} & \text{Pag2} & \ldots & \text{Pag 3.000.000.000} \end{array} \\ \left(\begin{array}{cccc} 0 & 1 & \ldots\ldots\ldots\ldots & 0 \\ . & . & \ldots\ldots\ldots\ldots & . \\ . & . & \ldots\ldots\ldots\ldots & . \\ 1 & . & \ldots\ldots\ldots\ldots & . \\ . & . & \ldots\ldots\ldots\ldots & . \\ . & . & \ldots\ldots\ldots\ldots & . \\ . & . & \ldots\ldots\ldots\ldots & 1 \\ . & . & \ldots\ldots\ldots\ldots & . \\ . & . & \ldots\ldots\ldots\ldots & . \\ 0 & 0 & \ldots\ldots\ldots\ldots & . \end{array} \right) \end{array}$$

Demasiada Informacion!

- Bueno: se puede paralelizar!

Demasiada Informacion!

- Bueno: se puede paralelizar!
- Malo: hay que almacenar y mover mucha informacion!

Demasiada Informacion!

- Bueno: se puede paralelizar!
- Malo: hay que almacenar y mover mucha informacion!
- Bit 0 o 1 minima informacion posible.

Demasiada Informacion!

- Bueno: se puede paralelizar!
- Malo: hay que almacenar y mover mucha informacion!
- Bit 0 o 1 minima informacion posible.
- Byte son 8 bits: p.ej. 10101110

Demasiada Informacion!

- Bueno: se puede paralelizar!
- Malo: hay que almacenar y mover mucha informacion!
- Bit 0 o 1 minima informacion posible.
- Byte son 8 bits: p.ej. 10101110
- Almacenar un numero en una PC ocupa 4 bytes (precisión simple)

Demasiada Informacion!

- Bueno: se puede paralelizar!
- Malo: hay que almacenar y mover mucha informacion!
- Bit 0 o 1 minima informacion posible.
- Byte son 8 bits: p.ej. 10101110
- Almacenar un numero en una PC ocupa 4 bytes (precisión simple)
- Informacion en la matriz:

$$300.000 \times 3.000.000.000 = 90.000.000.000.000 \sim 360.000\text{Gigas}$$

Demasiada Informacion!

- Bueno: se puede paralelizar!
- Malo: hay que almacenar y mover mucha informacion!
- Bit 0 o 1 minima informacion posible.
- Byte son 8 bits: p.ej. 10101110
- Almacenar un numero en una PC ocupa 4 bytes (precisión simple)
- Informacion en la matriz:

$$300.000 \times 3.000.000.000 = 90.000.000.000.000 \sim 360.000\text{Gigas}$$

- Con discos de 100 Gigas ... 3600 discos rígidos!!!

La matematica vuelve a ayudar!

Se usan técnicas para "achicar" la información.
LSI (Indexado por Semantica Latente)
Descomposicion en valores singulares:

$$G = UDV$$

No hace falta tomar toda la matriz D para "aproximar" G. Para saber mas de este tema:

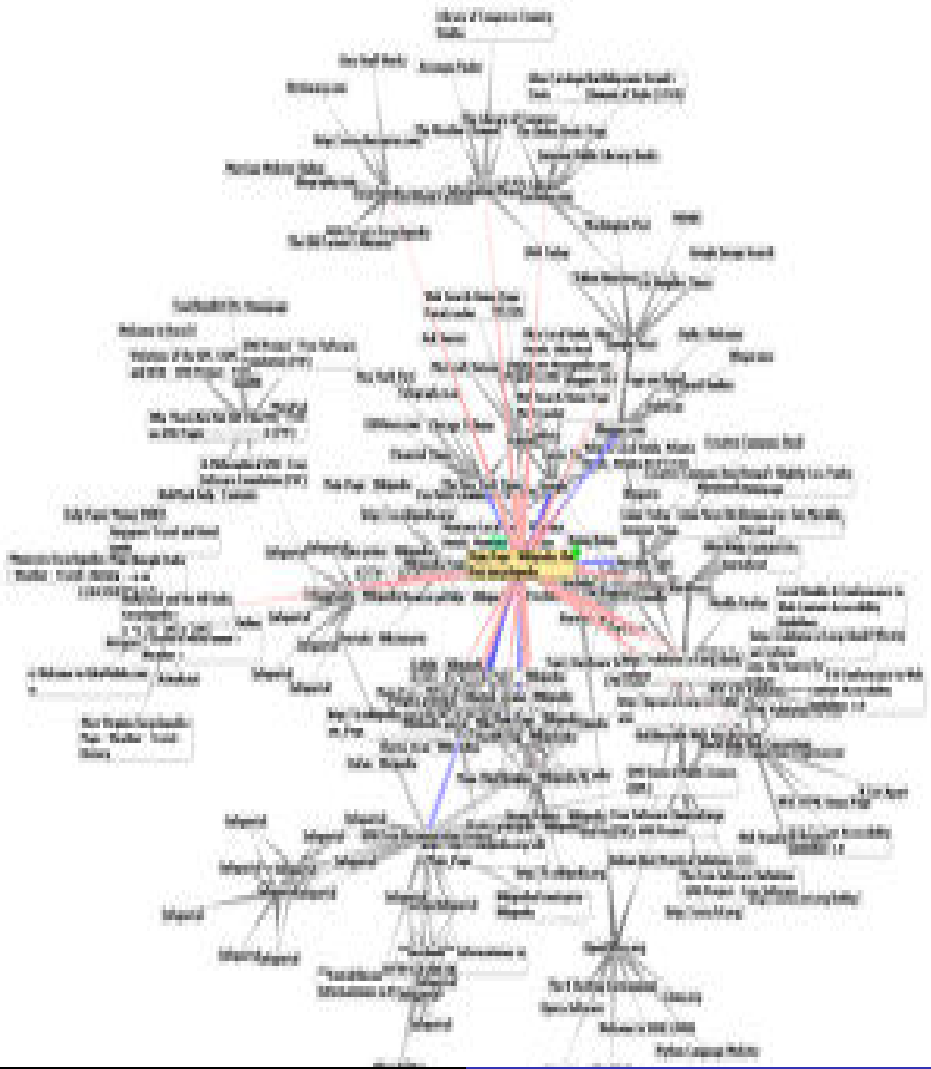
La matematica vuelve a ayudar!

Se usan técnicas para "achicar" la información.
LSI (Indexado por Semantica Latente)
Descomposicion en valores singulares:

$$G = UDV$$

No hace falta tomar toda la matriz D para "aproximar" G. Para saber mas de este tema: Pueden hacer la Licenciatura en Matemática!

El Problema



Criterio de Importancia de Paginas: el PageRank

- www.uam.es/personal_pdi/ciencias/gallardo/upm_google.pdf

Criterio de Importancia de Paginas: el PageRank

- www.uam.es/personal_pdi/ciencias/gallardo/upm_google.pdf
- Como cuantificar la importancia de una pagina?

Criterio de Importancia de Paginas: el PageRank

- www.uam.es/personal_pdi/ciencias/gallardo/upm_google.pdf
- Como cuantificar la importancia de una pagina?
- Una vez mas consideremos todas la paginas de Internet:

$$P_1, P_2, ..., P_{3.000.000.000}$$

y llamemos

$$x_1, x_2, ..., x_{3.000.000.000}$$

a la importancia que le atribuimos a cada una.

Criterio de Importancia de Paginas: el PageRank

- www.uam.es/personal_pdi/ciencias/gallardo/upm_google.pdf
- Como cuantificar la importancia de una pagina?
- Una vez mas consideremos todas la paginas de Internet:

$$P_1, P_2, ..., P_{3.000.000.000}$$

y llamemos

$$x_1, x_2, ..., x_{3.000.000.000}$$

a la importancia que le atribuimos a cada una.

- Google utiliza un algoritmo llamado PageRank

Mirando "Links"

El "truco" para cuantificar la importancia de cada pagina esta en relacionarla con las demás: esto se hace viendo los links que recibe.

- Si una pagina recibe muchos links debe ser importante

Mirando "Links"

El "truco" para cuantificar la importancia de cada pagina esta en relacionarla con las demás: esto se hace viendo los links que recibe.

- Si una pagina recibe muchos links debe ser importante
- Pero si yo recibo un solo link proveniente de Microsoft?????

Mirando "Links"

El "truco" para cuantificar la importancia de cada pagina esta en relacionarla con las demás: esto se hace viendo los links que recibe.

- Si una pagina recibe muchos links debe ser importante
- Pero si yo recibo un solo link proveniente de Microsoft?????
- La importancia de una pagina es proporcional a la suma de las importancias de las paginas que linkean a ella.

Mas Matemática!

- $x_1 = K(x_{27} + x_{1235})$

Mas Matemática!

- $x_1 = K(x_{27} + x_{1235})$
- Viejo conocido: Sistema de ecuaciones!!!

$$\begin{cases} x_1 = K(x_{27} + x_{1235}) \\ x_2 = K(x_{132} + x_{1256} + x_{5689}) \\ . = ... \\ . = ... \\ . = ... \\ x_{3.000.000.000} = K(x_{45} + x_{67}) \end{cases}$$

Mas Matemática!

- $x_1 = K(x_{27} + x_{1235})$
- Viejo conocido: Sistema de ecuaciones!!!

$$\begin{cases} x_1 = K(x_{27} + x_{1235}) \\ x_2 = K(x_{132} + x_{1256} + x_{5689}) \\ . = ... \\ . = ... \\ . = ... \\ x_{3.000.000.000} = K(x_{45} + x_{67}) \end{cases}$$

- 3.000.000.000 de ecuaciones y 3.000.000.001 incognitas!!!

Problemas de autovalores

- Se llama un problema de autovalores

Problemas de autovalores

- Se llama un problema de autovalores
- Aplicaciones en Ingenieria, Fisica, Biologia, Economía

Problemas de autovalores

- Se llama un problema de autovalores
- Aplicaciones en Ingenieria, Fisica, Biologia, Economía
- Mucha teoría sobre esto: y no es nueva ...

Problemas de autovalores

- Se llama un problema de autovalores
- Aplicaciones en Ingenieria, Fisica, Biologia, Economía
- Mucha teoría sobre esto: y no es nueva ...
- $\sim$ 1910 Perron y Frobenius

Problemas de autovalores

- Se llama un problema de autovalores
- Aplicaciones en Ingenieria, Fisica, Biologia, Economía
- Mucha teoría sobre esto: y no es nueva ...
- $\sim$ 1910 Perron y Frobenius
- Para saber mas ...

Problemas de autovalores

- Se llama un problema de autovalores
- Aplicaciones en Ingenieria, Fisica, Biologia, Economía
- Mucha teoría sobre esto: y no es nueva ...
- $\sim$ 1910 Perron y Frobenius
- Para saber mas ...
- Licenciatura en Matemática!!!

Terminando!

- En 1939 Edward Kasner 10^{100}

Terminando!

- En 1939 Edward Kasner 10^{100}
- $$10^{100} = 1\underbrace{000000000000000...0}_{100 \text{ ceros}}$$

Terminando!

- En 1939 Edward Kasner 10^{100}
- $$10^{100} = 1\underbrace{000000000000000...0}_{100 \text{ ceros}}$$
- Googol

Terminando!

- En 1939 Edward Kasner 10^{100}
- $$10^{100} = 1\underbrace{000000000000000...0}_{100 \text{ ceros}}$$
- Googol
- De Arena y de Atomos.

Terminando!

- En 1939 Edward Kasner 10^{100}
- $$10^{100} = 1\underbrace{000000000000000...0}_{100 \text{ ceros}}$$
- Googol
- De Arena y de Atomos.
- En 1998 Sergei Brin y Lawrence Page crearon Google .