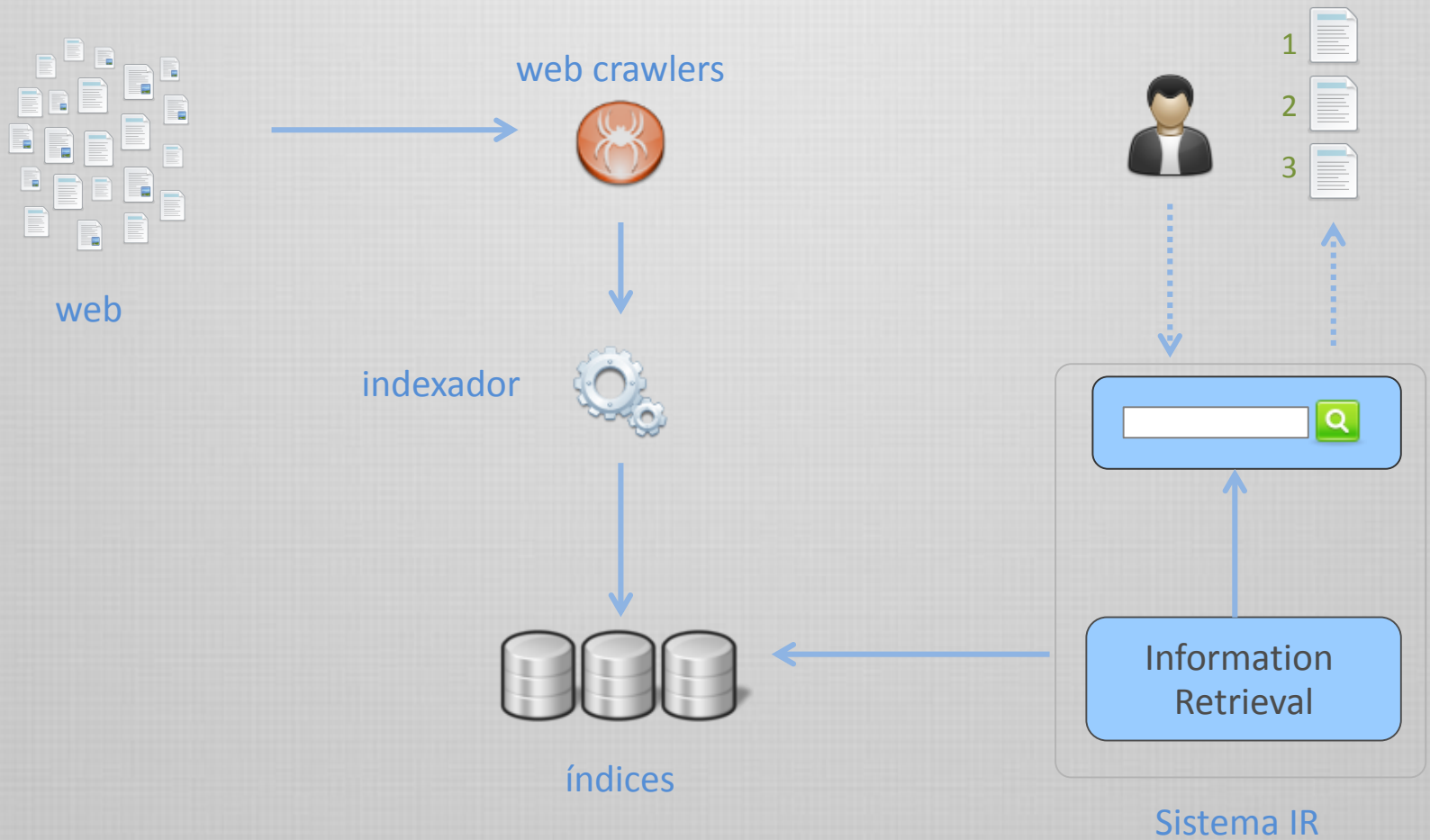


# Ingeniería de Aplicaciones Web

*Diego C. Martínez*

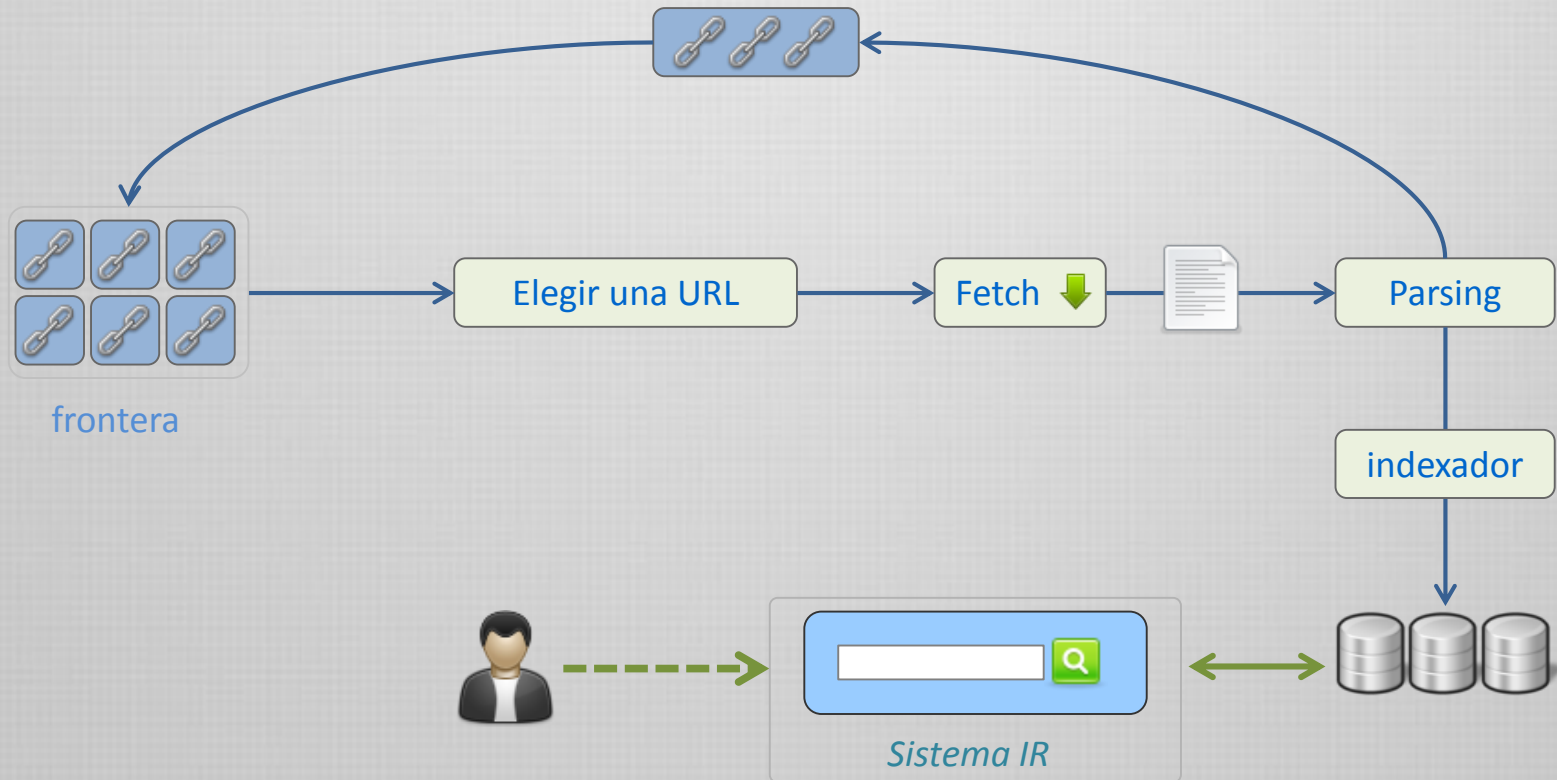
Departamento de Ciencias e Ingeniería de la Computación  
Universidad Nacional del Sur

# Buscadores web

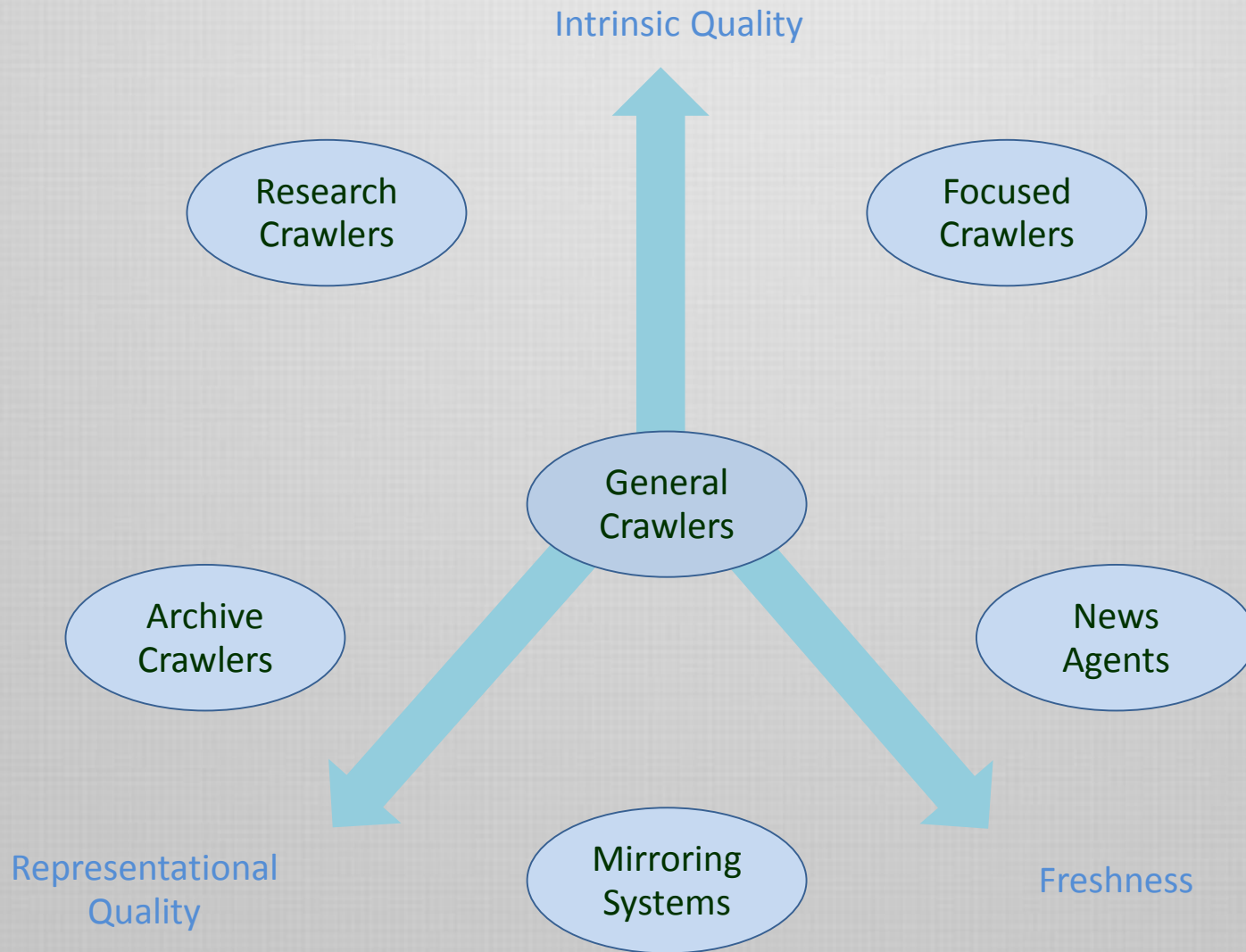


Actualmente el índice de Google  
ocupa más de 100 millones de gigabytes

# Web crawlers



# Taxonomia de web crawlers (Carlos Castillo)



# Relevancia

La determinación de la relevancia puede considerar varios aspectos:

- *Que el documento sea reciente*
- *Que el documento sea de fuentes confiables*
- *Que el documento satisfaga la necesidad del usuario*

Una noción simple de relevancia en búsquedas por palabras clave:

**Que las palabras clave figuren en el documento**

Una noción levemente exigente de relevancia en búsquedas por palabras clave:

**Que las palabras clave figuren en el documento *frecuentemente***

## Problemas con las palabras clave:

- Pueden no recuperar documentos con sinónimos  
*Por ejemplo, "PRC" vs. "China"*
- Pueden recuperar documentos irrelevantes  
*Por ejemplo, "Apple", "bat", "Médanos", "Paris Hilton"*
- Los términos correctos pueden ser desconocidos para el usuario




# Relevancia


El concepto predominante en la determinación de la relevancia hoy es TF-IDF

Term Frequency - Inverse Document Frequency

La frecuencia de un  
término en el documento



La frecuencia del término  
en el corpus



Alto ranking para palabras **poco frecuentes** que aparecen **mucho** en un documento

*Se buscan los documentos que contengan las palabras claves indicadas*

*Se calcula el TF y el IDF para cada término*

*Se suman las multiplicaciones de  $TF \cdot IDF$  de cada término*

El arte de la popularidad en la web: Search engine optimization (SEO)



# Relevancia

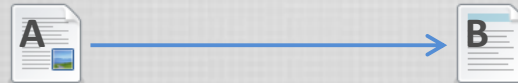
Algunas técnicas para mejorar la determinación de la relevancia:

- Tener en cuenta el significado de las palabras  
*Puede ayudar a descartar o priorizar documentos según el contexto*
- Tener en cuenta el orden de las palabras en la consulta  
*El orden puede esconder una prioridad mental del usuario con respecto a los términos*
- Registrar las reacciones de usuarios previos  
*¿Qué documentos eligieron otros usuarios que realizaron la misma búsqueda?*
- Extender la búsqueda a términos relacionados  
*No sólo sinónimos, sino también palabras cercanas en significado (“democracia”, “constitucion”, “república”, “libertad”)*
- Realizar correcciones ortográficas automáticas  
*Es sorprendente la cantidad de errores que cometen los usuarios.*
- Tener en cuenta la autoridad de la fuente.  
*Un artículo de la OMS sobre alguna enfermedad tendrá mayor relevancia que una opinión en un foro de discusión*
- Tener en cuenta la estructura de la URL  
*Las páginas con dominio “gov” serán probablemente más apropiadas para búsquedas como “ministerio” o “cancillería”*
- Tener en cuenta los links presentes en el documento  
*Son los “vecinos” de la página (“topic locality”)*

# Métricas de links

**PageRank** es el método de determinación de relevancia utilizado por Google y creado por Larry **Page** y Sergey Brin en 1995-98.

PageRank determina un valor numérico de importancia de una página determinada en la red, utilizando conceptos **estructurales y probabilísticos**.



Google interpreta un link desde la página A hacia la página B como un **voto** del autor de A a la página B

*En su forma más simple, una página tendrá mayor relevancia (ranking) si posee más votos.*





# Métricas de links

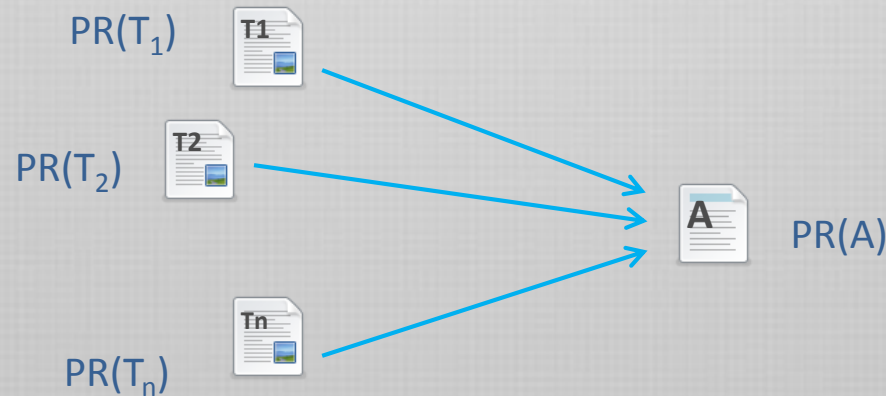
Google distingue también la **importancia** del voto

Una página con un voto importante tendrá mejor ranking que una con dos votos poco importantes

El cálculo de la importancia se basa en la estructura del documento y es independiente de las consultas.

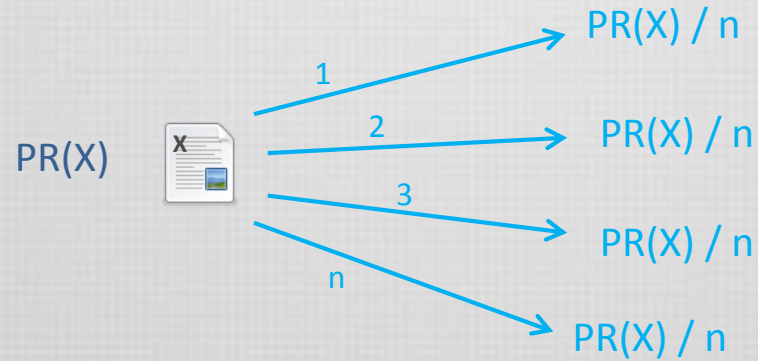
Incluye mas de 200 factores que influyen en la relevancia

La forma que tiene Google de calcular el PageRank es un secreto industrial, pero la estructura general es conocida...



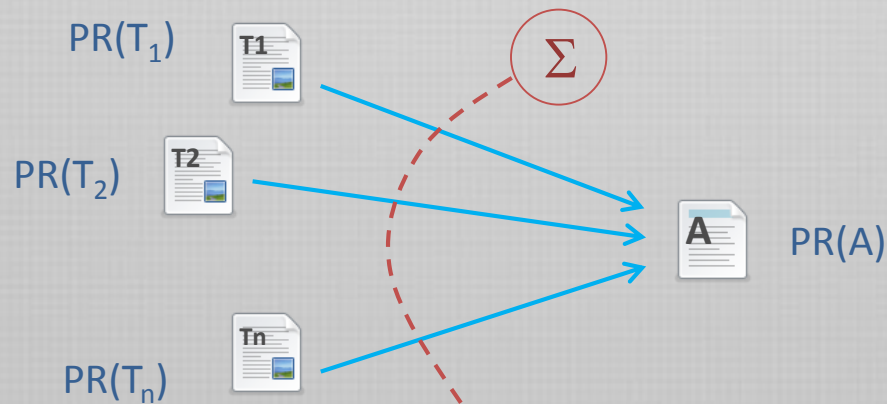
# Google PageRank

Una página *distribuye* su PageRank entre sus links de salida...



Cuantos más votos *reparte* una página, menos valen.

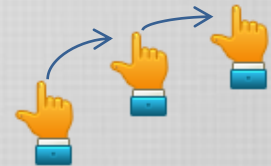
Los votos que una página recibe son acumulativos...



# Google PageRank

PageRank incluye una dimensión probabilística, basada en el *Modelo de Surfer*.

- La probabilidad de que el usuario haga click en un link está dada por la cantidad de links.
- La probabilidad de que un usuario **al azar** acceda una página, está dada por la suma de las probabilidades de los links que llevan a esa página.
- La probabilidad se **reduce** por un factor **d**, basado en el hecho de que el usuario no clickea infinitamente, sino que en algún momento **se aburre**.

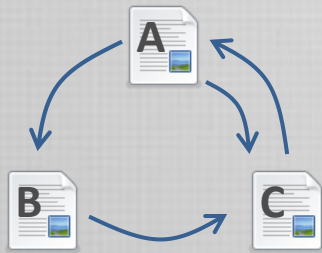


$$PR(A) = (1-d) + d ( PR(T_1) / C(T_1) + \dots + PR(T_n) / C(T_n) )$$

# Google PageRank

La web consiste de billones de documentos, por lo que en la práctica es necesario aplicar **algoritmos iterativos de aproximación** al valor de PageRank,

Se asignan valores iniciales de probabilidad y se iteran los cálculos del PageRank de cada página.  
*De acuerdo a Lawrence Page y Sergey Brin, aproximadamente 100 iteraciones son suficientes para conseguir buenos valores convergentes del PageRank.*



$d=0,5$

$$PR(A) = 0.5 + 0.5 PR(C)$$

$$PR(B) = 0.5 + 0.5 (PR(A) / 2)$$

$$PR(C) = 0.5 + 0.5 (PR(A) / 2 + PR(B))$$

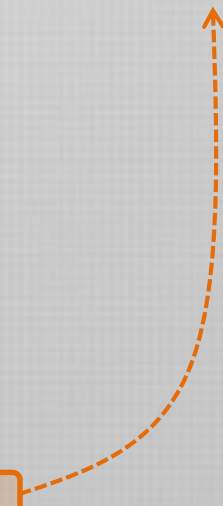


$$PR(A) = 1.07692308$$

$$PR(B) = 0.76923077$$

$$PR(C) = 1.15384615$$

Iteración	PR(A)	PR(B)	PR(C)
0	1	1	1
1	1	0.75	1.125
2	1.0625	0.765625	1.1484375
3	1.07421875	0.76855469	1.15283203
4	1.07641602	0.76910400	1.15365601
5	1.07682800	0.76920700	1.15381050
6	1.07690525	0.76922631	1.15383947
7	1.07691973	0.76922993	1.15384490
8	1.07692245	0.76923061	1.15384592



# Google PageRank

Actualmente Google combina el PageRank con otros factores de evaluación de relevancia.

Tips que se suelen mencionar para elevar el PageRank:

- ✓ *Agregar páginas nuevas al sitio.*
- ✓ *Intercambiar links con páginas con alto PageRank.*
- ✓ *Incrementar el número de in-links (publicidad)*

## Search Engine Optimization

“Optimización” de las páginas para obtener un ranking alto

### Acciones básicas

Title Tag  
Anchor Text  
Meta Keywords  
Meta Description  
Tips de Accesibilidad  
URLs cortas y descriptivas  
Navegación ordenada en el sitio  
Sitemaps  
...



# TrustRank



Parte del problema del tamaño de la web es el SPAM.



Clarks  
20% off  
Adult Styles

Superdry  
10% OFF  
First Time Orders when you sign up

WIN a Virgin Hot Air Balloon Package

LA fitness  
We'll get there together.  
FREE 3 day gym membership.  
Try everything and don't pay a penny.

A monster deal  
£50 M&S Voucher & FREE setup with HD orders

SUMMER SALE  
UP TO 50% OFF

Cath Kidston  
WIN a Cath Kidston Bunch Flowers Day Bag

a Virgin Hot Air Balloon Package for 2!

WIN A 50 VOUCHER  
spend at benefitcosmetics.co.uk

benefitcosmetics.co.uk

WIN A 100 VOUCHER

BEST BUY

SNAP UP A GREAT DEAL

7.5% OFF  
all Digital

WINNER  
YOU ARE THE 1,000,000th VISITOR  
YOU HAVE WON A FREE\* LAPTOP!  
CLICK HERE TO CLAIM!  
Claim number: 33AF3T

# TrustRank

Palabras más frecuentes de sitios SPAM



También

!!!

\$\$\$

100% free

# TrustRank

Existe desde hace muchos años una constante lucha para capturar la atención de los usuarios.

*Las técnicas generales de los buscadores son conocidas y algunos sitios **procuran engañarlas** para subir en el ranking*



**SPAMDEXING**

*Parte “sucia” de Search Engine Optimization*

*Es uno de los principales problemas de los motores de búsqueda  
Es necesario “descubrir” los sitios spam y bajarlos en el ranking*

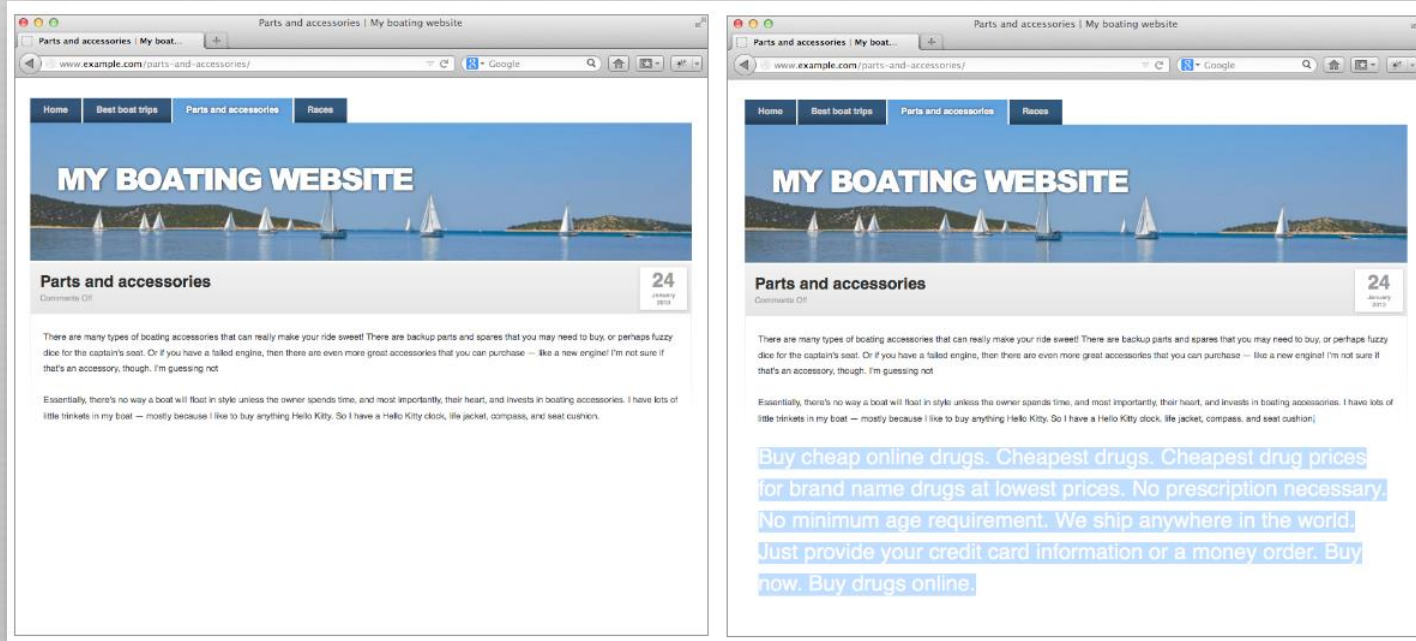


# TrustRank

Técnicas habituales de spam, ahora obsoletas:

## Texto Oculto

*El mismo color de font que de background*



## Link Farms

*Crear páginas falsas con links a una página puntual para incrementar su ranking.*

[Replica Watch](#) Exactwatches.com specializes in providing the Best (IN: 0 | OUT: 0)  
[Online Snowboard Shop](#) BoardersMall.com is a premier and leading company (IN: 0 | OUT: 0)  
[Plastic Recyclers](#) Vikoz Enterprises Waste Plastic Recycling Company (IN: 0 | OUT: 0)  
[Discount Granite Counter Tops](#) Stone By Nature is a premier and leading company w (IN: 0 | OUT: 0)  
[Indianapolis Wedding Bands](#) Welcome to Cool Chillies, here you will find inform (IN: 0 | OUT: 0)  
[HGVCub At Kalia Tower](#) All-Islands Timeshares is your source for the best (IN: 0 | OUT: 0)  
[Bed Comforters](#) Naturaworld.com is dedicated to helping you get th (IN: 0 | OUT: 0)  
[Machinery Buy & Sell](#) surplus, used business, industrial mach (IN: 0 | OUT: 0)  
[AquaBot Repair](#) AquaQualityPools Specialized in AquaBot Turbo Pool (IN: 0 | OUT: 0)  
[Electronic Pest Control](#) Ecolatermite.com, is well-known for providing Alte (IN: 0 | OUT: 0)  
[Dolan Designs](#) Lighting, Buy online and save on our large selecti (IN: 0 | OUT: 0)  
[Restaurant Loan](#) Online Information and resources for Creative Smal (IN: 0 | OUT: 0)  
[Divorce Lawyer Las Vegas](#) Gtogata.com web site is the only exclusive Las Veg (IN: 0 | OUT: 0)  
[Certainteed Roofing](#) Welcome to MGAroofing.com, here you will find info (IN: 0 | OUT: 0)  
[Exercise Weight Loss Program](#) Rhinomanprod.com is your online source for the Bes (IN: 0 | OUT: 0)  
[Bosley Alternative](#) Visit Samson Hair Restoration's web site for their (IN: 0 | OUT: 0)  
[Phoenix DUI Lawyer](#) At the Maasen Law Firm, you will find the highest (IN: 0 | OUT: 0)



# TrustRank

Técnicas habituales de spam, ahora obsoletas:

## Honey Pot

*Páginas con contenido real pero poblada de links a otros sitios spam*

## Keyword Abuse

*Registrar muchas palabras clave relevantes para los buscadores*

## Scraper Sites

*Sitios armados con información de motores de búsqueda y otros sitios*

## Article Spinning

*Copiar y (pegar + modificar) contenido*

## Cloaking

*Ofrecer páginas especiales al crawler*

Es necesario procurar descubrir estos sitios spam para no incluirlos en el ranking.

- No siempre son sitios falsos o ilegales.
- Ha ocurrido que el scraper de un sitio supera en ranking al sitio original.

Técnicas de detección: *machine learning, estudio de topología web, análisis de links, etc*

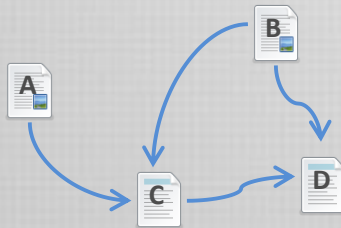
# TrustRank

TrustRank es una técnica algorítmica basada en PageRank.

*Toma nodos rankeados por PageRank y analiza el grado de relación de cada página con una página confiable.*

El score de cada página se calcula iterativamente ( $n$  iteraciones)

Utiliza el mismo modelo web que PageRank



- *in-links* = hiperlinks hacia una página  
*in-degree* = cantidad de in-links
- *out-links* = hiperlinks desde una página  
*out-degree* = cantidad de out-links

$$T(p, q) = \begin{cases} 0 & \text{if } (q, p) \notin \varepsilon \\ 1/\omega(q) & \text{if } (q, p) \in \varepsilon \end{cases}$$

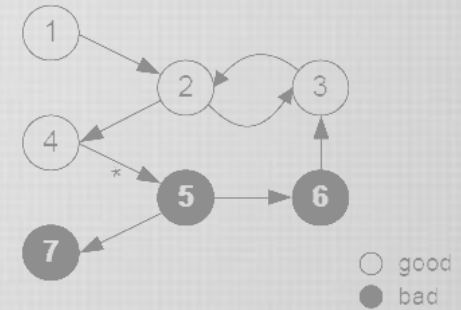
$$T = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

*"Combating Web Spam with TrustRank"* - Zoltan Gyongyi, Hector Garcia-Molina, Jan Pedersen

# TrustRank

Se formaliza la intervención humana con una función **oráculo**

$$O(p) = \begin{cases} 0 & \text{if } p \text{ es spam,} \\ 1 & \text{if } p \text{ es buena página.} \end{cases}$$



Se busca minimizar la invocación al oráculo.

Se aplica solo a un conjunto de páginas.

Se toma el principio de *approximate isolation*

*“buenas páginas rara vez apuntan a malas páginas”*

Se define  $T(p)$ : la función de confianza de una página  $p$

Idealmente:

$$T(p) = \Pr[O(p) = 1].$$

Ej: 100 paginas, cada una con Trust de 0.7  
*Con  $T$  ideal, el oráculo marcará 70 de esas páginas como 1,  
el resto como 0*

# TrustRank

Naturalmente es muy difícil conseguir la función ideal.  
Se definen parámetros de evaluación. Por ejemplo,

$$T(p) > \delta \Leftrightarrow O(p) = 1$$

Si la página recibe un puntaje por encima de  $\delta$ , entonces es confiable.

Es importante elegir bien el conjunto inicial de páginas confiables (seed set)

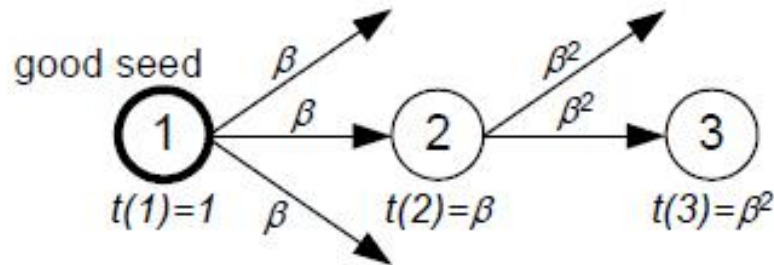
*Aleatorio*

*Inverse PageRank (las que tienen mas outlinks)*

*PageRank alto*

Una vez elegido el seed, se establece la propagación y atenuación de la confianza.

# TrustRank – atenuación de confianza



$$\beta < 1$$

Figure 3: Trust dampening.

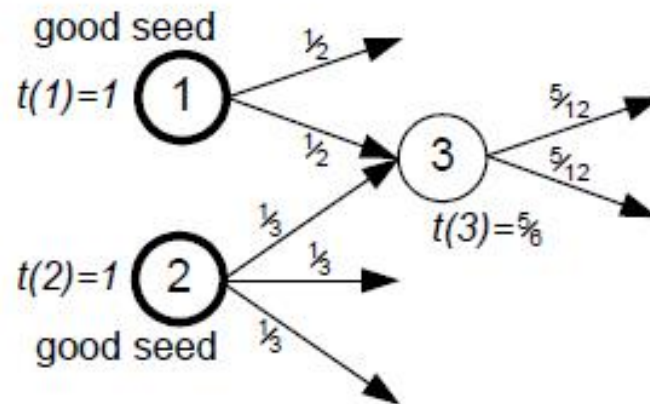
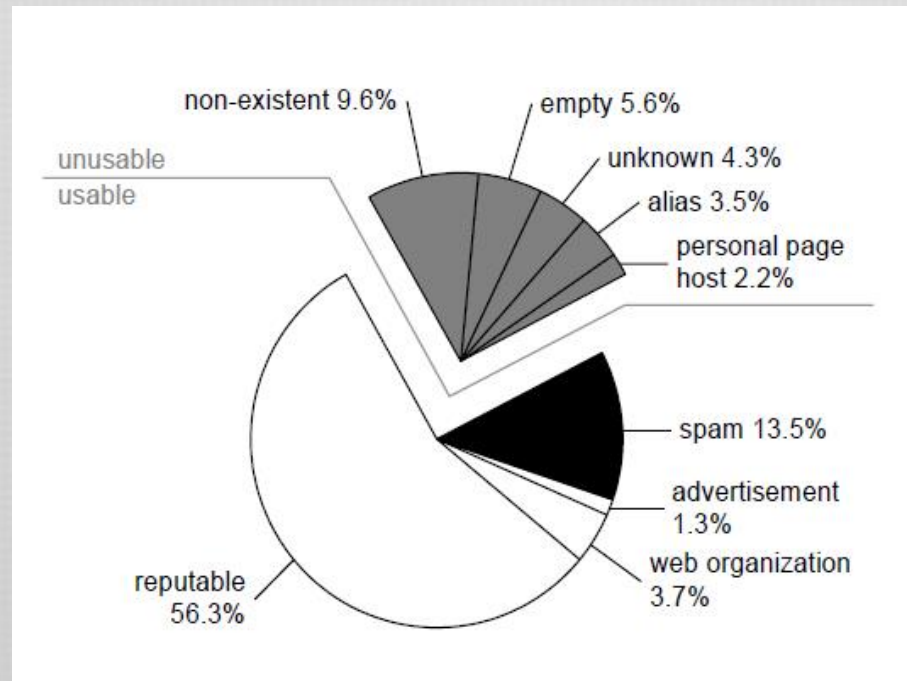


Figure 4: Trust splitting.



# TrustRank



Conjunto inicial de muestra

# TrustRank – resultados tests

