¿Son adecuados los buscadores modernos?

Estudio realizado sobre *logs* de *AlltheWeb*. Jansen, B.J. y Spink, A. 2003, "An Analysis of Web Documents Retrieved and Viewed", *The 4th International Conference on Internet Computing*, pp. 65-69.

- 24 horas
- 150.000 sesiones de usuario
- 450.000 consultas
- 13% de las consultas eran únicas
- 53% de las sesiones constituidas por una única consulta
- 54% de las sesiones sólo examinaron primera página de resultados
- 54% de las consultas sólo examinaron un único resultado
- 66% de las sesiones examinaron de 1 a 5 resultados
- Para 530 consultas evaluadas "manualmente" en el 48,5% de los casos el resultado visitado no era relevante

Sistemas y servicios informáticos para Internet (2007/08)
Oviedo, 3, 4 y 5 de Marzo de 2008
Departamento de Informática
Web Semántica

¿Son adecuados los buscadores modernos?

Estudio realizado sobre *logs* de *AlltheWeb*. Jansen, B.J. y Spink, A. 2003, "An Analysis of Web Documents Retrieved and Viewed", *The 4th International Conference on Internet Computing*, pp. 65-69.

Conclusiones Jansen y Spink

- Mayoría de usuarios tienen necesidades de información simples
- Los buscadores resuelven bien este tipo de consultas
- Usuario promedio necesita ver 2 documentos para encontrar 1 relevante

¿Mis conclusiones?

- Echémosle un ojo a este artículo que trabaja sobre los mismos *logs*…
- Jansen, B.J. y Spink, A. 2006, "How are we searching the World Wide Web? A comparison of nine search engine transaction logs", *Information Processing and Management*

¿Son adecuados los buscadores modernos?

Distribución temática de las consultas (2.503 consultas)

People, places or things	22,50%
Computers or Internet	21,80%
Commerce, travel, employment, or economy	12,30%
Entertainment or recreation	9,10%
Sex and pornography	10,80%
Health or sciences	7,80%
Society, culture, ethnicity, or religion	4,80%
Performing or fine arts	4,70%
Education or humanities	2,90%
Government or legal	2,70%
Non–English or unknown	0,60%

42,4% de todas las consultas son sobre famosos, ocio y sexo ("fáciles")

55%-84% de las consultas más frecuentes son análogas (dependiendo de la lista el porcentaje de sexo varía entre el 3%, el 48% o el 60%)

Las consultas frecuentes suponen entre el 2% y el 18% del total de consultas

Este tipo de consultas constituyen el 45% del total…

Sistemas y servicios informáticos para Internet (2007/08)
Oviedo, 3, 4 y 5 de Marzo de 2008
Departamento de Informática
Web Semántica

¿Son adecuados los buscadores modernos?

Más datos (elaboración propia)

Relevancia promedio de los resultados está, efectivamente, alrededor del 50%

Sin embargo, ¿cuál es la dispersión?

20% consultas de la muestra tiene una precisión media del 21%

23% consultas no obtienen ningún resultado relevante en la primera página

Estimación: 15%-20% todas las consultas no obtienen resultados relevantes

Mis conclusiones

Casi la mitad de las consultas son relativas a famosos, ocio y sexo (es decir, "fáciles" de sastisfacer)

En consecuencia, casi la mitad de los usuarios quedan satisfechos con los resultados

Pero… Un porcentaje sustancial de consultas exige a los usuarios "bucear" más allá de la primera página de resultados

(Más) Problemas del *ranking* basado en hiperenlaces

Tres escenarios problemáticos

- Bharat, K., y Henzinger, M. 1998, "Improved Algorithms for Topic Distillation in a Hyperlinked Environment", en *Proceedings of SIGIR-98, 21st ACM International Conference on Research and Development in Information Retrieval*, pp. 104-111.

Enlaces "nepotistas"

Cada enlace es un "voto" aunque provengan todos del mismo servidor

No es un problema fácil de resolver. Davison, B.D. 2000, "Recognizing Nepotistic Links on the Web", en *Proceedings of AAAI-2000 Workshop on Artificial Intelligence for Web Search*, pp. 23-28.

Enlaces automáticos

Todos estos algoritmos parten del supuesto que los enlaces son establecidos por un ser humano y eso no siempre es cierto (*Wordpress scandal*)

Documentos irrelevantes enlazados desde autoridades

Inevitable puesto que no hay ningún análisis de contenidos, sólo se emplea la topología del grafo

Sistemas y servicios informáticos para Internet (2007/08)
Oviedo, 3, 4 y 5 de Marzo de 2008
Departamento de Informática
Web Semántica

(Más) Problemas del *ranking* basado en hiperenlaces

Page, L., Brin, S., Motwani, R. y Winograd, T. 1998, *The PageRank Citation Ranking: Bringing Order to the Web*

> *[…] PageRanks are virtually immune to manipulation by commercial interests. For a page to get a high PageRank, it must convince an important page, or a lot of non-important pages to link to it. At worst, you can have manipulation in the form of buying advertisements (links) on important sites. But, this seems well under control since it costs money.*

(Más) Problemas del *ranking* basado en hiperenlaces

Page, L., Brin, S., Motwani, R. y Winograd, T. 1998, *The PageRank Citation Ranking: Bringing Order to the Web*

> *[…] PageRanks are virtually immune to manipulation by commercial interests. For a page to get a high PageRank, it must convince an important page, or a lot of non-important pages to link to it. At worst, you can have manipulation in the form of buying advertisements (links) on important sites. But, this seems well under control since it costs money.*

Sistemas y servicios informáticos para Internet (2007/08)
Oviedo, 3, 4 y 5 de Marzo de 2008
Departamento de Informática
Web Semántica

(Más) Problemas del *ranking* basado en hiperenlaces

Granjas de enlaces

Recordemos que el *PageRank* total es constante, sólo se reparte entre los nodos

¿Qué sucede si se construye un grafo con gran cantidad de nodos fuertemente conectados y unos pocos reciben la mayoría de enlaces?

Respuesta: Una porción del *PageRank* global termina en ese subgrafo y es asignada en su práctica totalidad a unos pocos nodos que aumentan su *PageRank* artificialmente

Si, además, la granja de enlaces es alojada o enlazada desde algún sitio "prestigioso" mejor que mejor

(Más) Problemas del *ranking* basado en hiperenlaces

Google bombing

Además de emplear la topología derivada de los enlaces *Google* (y el resto de buscadores) emplea/ba el texto de los enlaces que recibe una página para indexarla (ej. **`google compra youtube`**)

Si varios sitios web coordinados enlazan a un tercero empleando el mismo término o frase es posible construir "bromas" como las famosas: **`ladrones`**, **`miserable failure`** o **`horrid operating system`**

A finales de enero de 2007 *Google* anunció que las "bombas" ya no funcionaban gracias a una solución algorítmica

Pero…

- Algunas siguen funcionando: **`horrid operating system`**
- Búsquedas que deberían funcionar no lo hacen: **`spanish airlines`**
- Todos los buscadores son vulnerables a esta técnica

Off-topic: ¿qué retornan las consultas **`click here`** o **`pinche aquí`**?

Sistemas y servicios informáticos para Internet (2007/08)
Oviedo, 3, 4 y 5 de Marzo de 2008
Departamento de Informática
Web Semántica

(Más) Problemas del *ranking* basado en hiperenlaces

Daños "colaterales" (usuarios y autores)

La ausencia de "prestigio" no implica carencia de relevancia

Al desvincularse el "prestigio" de los contenidos, resultados "prestigiosos" pueden satisfacer la consulta pero no al usuario.

El autor del documento puede no desear tales visitas.

Algunas consultas reales que me han traído "público"…

```
algoritmos genéticos(documentos en inglés)
que es disertacion doctoral
articulos o comentarios de prensa sobre algoritmos geneticos
```

¿Qué tienen en común? Son consultas informativas, el usuario quiere información/respuestas. Broder, A. 2002, "A taxonomy of web search", *ACM SIGIR Forum*, vol. 36, no. 2, pp. 3-10. ★

(Más) Problemas del *ranking* basado en hiperenlaces

Web Spam Challenge
`http://webspam.lip6.fr/`

En la Web también hay *spam* **L** y es preciso detectarlo y luchar contra él… Por esa razón se habla de *adversarial information retrieval*

Algunos artículos interesantes:

Gyöngyi, Z. *et al.* 2004, "Combating web spam with TrustRank"
Gyöngyi, Z. y Garcia-Molina, H. 2005, "Web spam taxonomy"
Fetterly, D. *et al.* 2004, "Spam, damn spam, and statistics: using statistical analysis to locate spam web pages"
Benczúr, A.A. *et al.* 2005, "SpamRank–Fully Automatic Link Spam Detection Work in progress"
Ntoulas, A. *et al.* 2006, "Detecting spam web pages through content analysis"
Becchetti, L. *et al.* 2006, "Link-Based Characterization and Detection of Web Spam"
Castillo, C. *et al.* 2006, "A reference collection for web spam"

Sistemas y servicios informáticos para Internet (2007/08)
Oviedo, 3, 4 y 5 de Marzo de 2008
Departamento de Informática
Web Semántica

No hay talla única...

Los buscadores actuales son muy buenos…

…localizando sitios web conocidos/"oficiales"
…facilitando el acceso a servicios *on-line* (mapas, tiempo, *e-mail*, subastas, etc.)
…resolviendo consultas simples (famosos, ocio y sexo)

En suma… Satisfaciendo a la mayor parte de la gente la mayor parte del tiempo

Pero como fuente de información la Web sigue siendo…

... y recordemos que las suposiciones son falsas

La Web no es un grafo fuertemente conectado,

Broder, A. et al. 2000, "Graph structure in the web: experiments and models", en *Proceedings of the ninth WWW Conference*

Sólo el 90% de la Web está fuertemente conectada

Meiss, M.R. *et al.* 2008, "Ranking web sites with real user traffic" ★

> *PageRank ranks sites very differently than actual human traffic, especially for the most important hosts. This finding is interpreted in light of our empirical analysis, showing how each of the random behavior assumptions underlying PageRank is violated: not all links from a site are followed equally, but even more importantly, some sites are much more likely than others to be the starting or ending points of surfing sessions.*

Presente y futuro de la Web

- ¿Web 2.0?
- Filtrado colaborativo
- Personalización
- Minería Web *(Web Mining)*
- Análisis de tendencias
- *Normalized Google Distance*
- La Web como *corpus*
- Para saber más…
- Evolución

¿Web 2.0?

O'Reilly, T. 2006, *Web 2.0 Compact Definition: Trying Again*

> *Web 2.0 is the business revolution in the computer industry caused by the move to the internet as platform, and an attempt to understand the rules for success on that new platform. Chief among those rules is this: Build applications that harness network effects to get better the more people use them.*

Sistemas y servicios informáticos para Internet (2007/08)
Oviedo, 3, 4 y 5 de Marzo de 2008
Departamento de Informática
Web Semántica

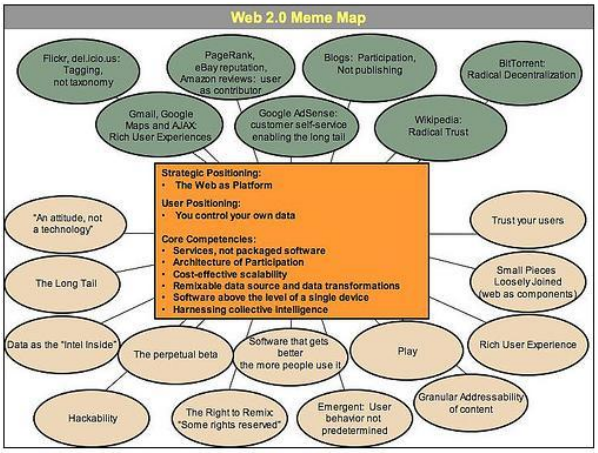
¿Web 2.0?

¿Son lo mismo Web 2.0 y Web Semántica?

Según Tim Berners-Lee se parecen lo mismo que un huevo a una castaña (*"chalk and cheese"*)... Sin embargo, son buenas por separado y mucho mejor juntas

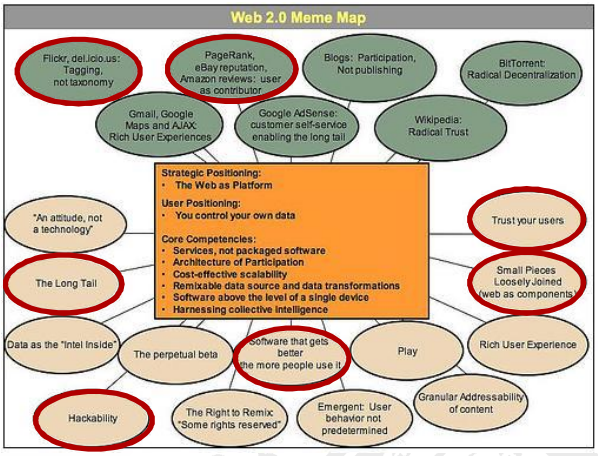
Según el *W3C* ambas se centran en el intercambio y combinación de datos heterogéneos pero la Web Semántica proporciona un modelo más consistente

¿Web 2.0?



Sistemas y servicios informáticos para Internet (2007/08)
Oviedo, 3, 4 y 5 de Marzo de 2008
Departamento de Informática
Web Semántica

¿Web 2.0?



¿Web 2.0?

También hay algunos puntos interesantes…

> *People subscribe to each others' sites, and easily link to individual comments on a page, but also, via […] trackbacks, they can see when anyone else links to their pages, and can respond […] Interestingly, two-way links were the goal of early hypertext systems like Xanadu. Hypertext purists have celebrated trackbacks as a step towards two way links.*

…

> *(The Long Tail) Small sites make up the bulk of the internet's content; […] Therefore: Leverage customer-self service and algorithmic data management to reach out to the entire web, to the edges and not just the center, to the long tail and not just the head.*

…

> *The key to competitive advantage in internet applications is the extent to which users add their own data to that which you provide.*

Sistemas y servicios informáticos para Internet (2007/08)
Oviedo, 3, 4 y 5 de Marzo de 2008
Departamento de Informática
Web Semántica

¿Web 2.0?

Folksonomía (*folksonomy = folk + taxonomy*)

Una forma de metadatos

Etiquetado colaborativo de recursos en la Web

Las etiquetas no proceden de un vocabulario controlado sino que son elegidas libremente por los usuarios

La co-ocurrencia de etiquetas lleva a las folksonomías…

Otra cuestión es cómo emplearlas (más allá de la búsqueda por "serendipia")

Ejemplos: *del.icio.us*, *flickr, tagzania*

Problemas: sinonimia, polisemia, acrónimos, términos multipalabra, multilingüismo…

Mathes, A. 2004, *Folksonomies – Cooperative Classification and Communication Through Shared Metadata*

Filtrado colaborativo

Un sistema de filtrado de información procesa grandes volúmenes de datos para transmitir al usuario sólo aquellos *items* con mayores probabilidades de ser "interesantes"

El filtrado puede hacerse en base al contenido de los *items* o en base al juicio de otros usuarios del sistema (colaborativo)

El filtrado colaborativo no es reciente…

★ Goldberg, D., Nichols, D., Oki, B.M. y Terry, D. 1992, "Using Collaborative Filtering to Weave an Information Tapestry", *Com. of the ACM*, vol.35, no.12, pp. 61-70.

…*Amazon* lleva usándolo desde hace bastante tiempo (*"Customers who bought this book also bought"*). Funciona muy bien porque los usuarios "votan con dólares"

Otros ejemplos: *last.fm* (música), *IMDB* (películas)

Sistemas y servicios informáticos para Internet (2007/08)
Oviedo, 3, 4 y 5 de Marzo de 2008
Departamento de Informática
Web Semántica

Personalización

No hay talla única. Cada usuario es un mundo…

Lo ideal sería darle a cada persona justo lo que necesita

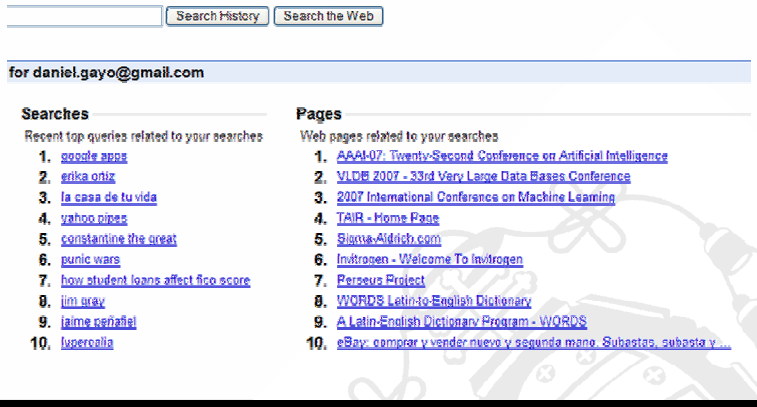
El problema es ¿cómo?

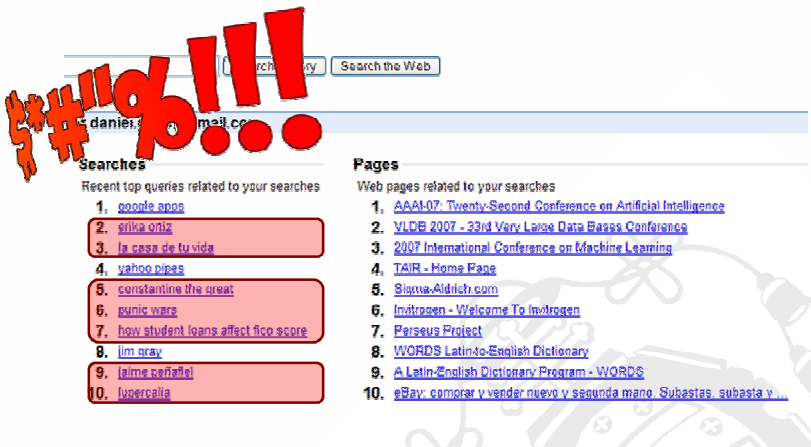
PageRank personalizado. Recordemos el modelo del *random surfer*, había una probabilidad *d* de "saltar" a una página cualquiera de la Web; sin embargo no todas las páginas de la Web tienen que ser equiprobables. Page, L., *et al.* 1998, *The PageRank Citation Ranking: Bringing Order to the Web*

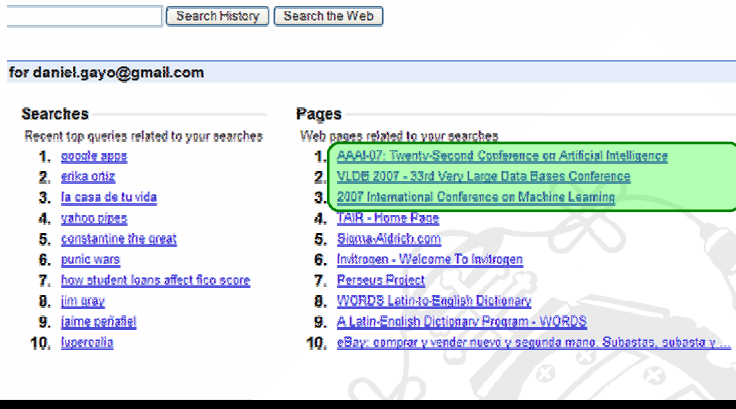
Inviable, no se puede calcular el *PageRank* para toda la Web y cada usuario

Otra posible solución radicaría en calcular el *PageRank* tras dividir la Web en subgrafos "temáticos"; después se personalizaría la consulta en base a la temática de la misma y/o la detectada en el contexto del usuario. Haveliwala, T.H. 2003, "Topic-sensitive pagerank: A context-sensitive ranking algorithm for web search", *IEEE Transactions on Knowledge and Data Engineering*

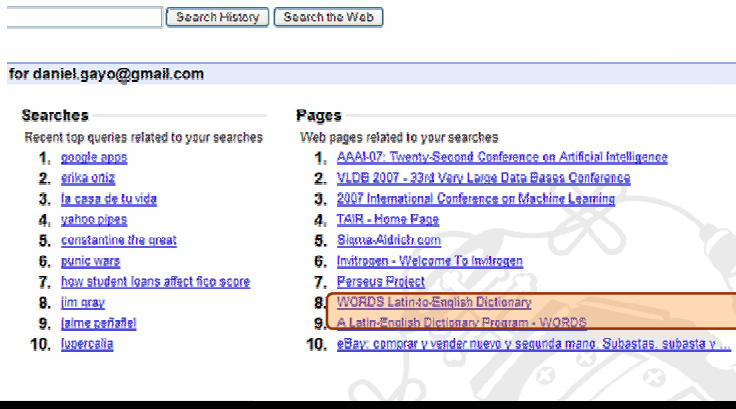
Explotando el historial de búsquedas y consultas. Lawrence, S. 2000, "Context in Web Search", *IEEE Data Engineering Bulletin*, vo. 23, no. 3, pp. 25–32

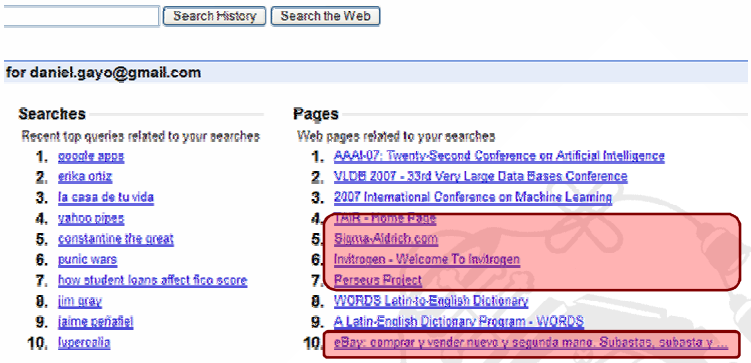

Search History
Search the Web
for daniel.gayo@gmail.com
Searches
Recent top queries related to your searches
1. google apps
2. erika ortiz
3. la casa de tu vida
4. yahoo pipes
5. constantine the great
6. punic wars
7. how student loans affect fico score
8. jim gray
9. jaime peñafiel
10. lupercalia
Pages
Web pages related to your searches
1. AAAI-07: Twenty-Second Conference on Artificial Intelligence
2. VLDB 2007 - 33rd Very Large Data Bases Conference
3. 2007 International Conference on Machine Learning
4. TAIR - Home Page
5. Sigma-Aldrich.com
6. Invitrogen - Welcome To Invitrogen
7. Perseus Project
8. WORDS Latin-to-English Dictionary
9. A Latin-English Dictionary Program - WORDS
10. eBay: comprar y vender nuevo y segunda mano. Subastas, subasta y ...


$#!@%!!!
Search the Web
Searches
Recent top queries related to your searches
1. google apps
2. erika ortiz
3. la casa de tu vida
4. yahoo pipes
5. constantine the great
6. punic wars
7. how student loans affect fico score
8. jim gray
9. jaime peñafiel
10. lupercalia
Pages
Web pages related to your searches
1. AAAI-07: Twenty-Second Conference on Artificial Intelligence
2. VLDB 2007 - 33rd Very Large Data Bases Conference
3. 2007 International Conference on Machine Learning
4. TAIR - Home Page
5. Sigma-Aldrich.com
6. Invitrogen - Welcome To Invitrogen
7. Perseus Project
8. WORDS Latin-to-English Dictionary
9. A Latin-English Dictionary Program - WORDS
10. eBay: comprar y vender nuevo y segunda mano. Subastas, subasta y ...



Sistemas y servicios informáticos para Internet (2007/08)
Oviedo, 3, 4 y 5 de Marzo de 2008
Departamento de Informática
Web Semántica





Sistemas y servicios informáticos para Internet (2007/08)
Oviedo, 3, 4 y 5 de Marzo de 2008
Departamento de Informática
Web Semántica

Minería web *(Web Mining)*

La extracción de conocimiento de la Web, minería Web o *Web mining* tiene como objetivo extraer información útil mediante el procesamiento de los ingentes volúmenes de datos que existen en la Web y que se generan con su uso diario

La minería Web puede dividirse en tres grandes áreas:

- Extracción de conocimiento a partir de la estructura hipertextual de la Web (p.ej. algoritmos *PageRank* y *HITS*)
- Extracción de conocimiento a partir del uso de la Web (p.ej. *logs* de servidores y buscadores)
- Extracción de conocimiento a partir de los contenidos disponibles en la Web (la Web como *corpus*)

Multidisciplinar: aprendizaje automático, procesamiento de lenguaje natural, estadística, recuperación de información, bases de datos

Minería web *(Web Mining)*

Los buscadores modernos son un ejemplo del conocimiento que se puede derivar de la estructura topológica de la Web

Los sistemas de filtrado colaborativo obtienen conocimiento a partir de las acciones de los usuarios en un sitio web concreto (podría considerarse un caso particular de minería de uso)

Aplicar minería de datos a los archivos de *log* de un servidor web no es nuevo

> Mobasher, B. *et al.* 1996, *Web Mining: Pattern Discovery from World Wide Web Transactions*, informe técnico, Universidad de Minnesota

Sistemas y servicios informáticos para Internet (2007/08)
Oviedo, 3, 4 y 5 de Marzo de 2008
Departamento de Informática
Web Semántica

Minería web *(Web Mining)*

Todos los servidores web generan archivos de *log* en los que se recoge información sobre las acciones de los usuarios en el sitio web

```
156.35.14.9 - - [17/Oct/2006:20:34:26 +0200] "GET /nol/shared/css/news_r5.css HTTP/1.0" 404 312
156.35.14.9 - - [17/Oct/2006:20:34:26 +0200] "GET /shared/css/toolbar_banner.css HTTP/1.0" 404 315
156.35.14.9 - - [17/Oct/2006:20:35:23 +0200] "GET /CursoWeb20/ HTTP/1.0" 200 1894
156.35.14.9 - - [17/Oct/2006:20:35:23 +0200] "GET /icons/blank.gif HTTP/1.0" 200 148
156.35.14.9 - - [17/Oct/2006:20:35:23 +0200] "GET /icons/back.gif HTTP/1.0" 200 216
156.35.14.9 - - [17/Oct/2006:20:35:23 +0200] "GET /icons/folder.gif HTTP/1.0" 200 225
156.35.14.9 - - [17/Oct/2006:20:35:23 +0200] "GET /icons/compressed.gif HTTP/1.0" 200 1038
```

En realidad, el servidor desconoce quién es el usuario, sólo dispone de su dirección IP que, en muchos casos, será utilizada por múltiples usuarios simultaneamente (*proxies*) y en otros será re-utilizada en diversas ocasiones (p.ej. direcciones dinámicas otorgadas por ISPs)

Por esa razón, lo máximo que puede hacerse con la información del archivo de *log* es tratar de encontrar sesiones de usuario (conjunto de peticiones realizadas desde una misma IP durante un período corto de tiempo)

Minería web *(Web Mining)*

Objetivos de la extracción de conocimiento a partir de archivos de *log*:

- Entender los intereses de los usuarios de un sitio web
- Mejorar, en consecuencia, la satisfacción del usuario al reorganizar el sitio en base a dichos intereses
- Facilitar el acceso a la información mediante recomendaciones en tiempo real

El último objetivo también puede alcanzarse empleando *swarm intelligence*

★ Wu, J. y Aberer, K. 2003, "Swarm Intelligent Surfing in the Web", *ICWE 2003*, LNCS 2722, pp. 431–440

Off-topic: video sobre robots, *swarm intelligence* y rastros de feromonas
`http://www.youtube.com/watch?v=z3E86D4dKN4`

Sistemas y servicios informáticos para Internet (2007/08)
Oviedo, 3, 4 y 5 de Marzo de 2008
Departamento de Informática
Web Semántica

Minería web *(Web Mining)*

Los motores de búsqueda también disponen de archivos de *log* en los que se almacena información como:

- Identificador de sesión
- Fecha y hora
- Texto de la consulta
- URL visitada
- Posición de la URL visitada dentro de la página de resultados

Minería web
(Web Mining)

Los *logs* de un buscador resultan muy útiles para mejorar la precisión

Baeza-Yates, R. 2004, "Query Usage Mining in Search Engines", en *Web Mining: Applications and Techniques*

> *After a query, a user usually performs a click to view one answer page. Each click is considered a positive recommendation of that page (in most cases bad pages are not clicked).*

★ Zhang, D. y Dong, Y. 2002, "A novel Web usage mining approach for search engines", *Computer Networks*, vol. 39, no. 3, pp. 303-310

> *A user is "good" if he/she issues many "good" queries, while a queryis "good" if it can retrieve many "good" resources, while a resource is "good" if it is accessed by many "good" users.*

Sistemas y servicios informáticos para Internet (2007/08)
Oviedo, 3, 4 y 5 de Marzo de 2008
Departamento de Informática
Web Semántica

Minería web
(Web Mining)

Los *logs* de un buscador resultan muy útiles para mejorar la precisión

Joachims, T. "Optimizing Search Engines Using Clickthrough Data", *Proceedings of the ACM Conference on Knowledge Discovery and Data Mining (KDD)*

`http://svmlight.joachims.org/`

Más artículos interesantes:

"Accurately Interpreting Clickthrough Data as Implicit Feedback"

"Query-Log Based Authority Analysis for Web Information Search"

"Optimizing Web Search using Spreading Activation on the Clickthrough Data"

Minería web *(Web Mining)*

Ricardo Baeza-Yates lleva algún tiempo desarrollando técnicas para agrupar consultas temáticamente a partir de los resultados visitados.

★ Baeza-Yates, R. *et al.* 2004, "Query recommendation using query logs in search engines", en *Current Trends in Database Technology*, LNCS 3268, p. 588-596.

Las consultas de un grupo pueden ordenarse en base al porcentaje de documentos relevantes (determinados por los *clicks*) que retorna cada una

Una vez determinados los conjuntos de consultas pueden emplearse para…

… ofrecer consultas alternativas

… mejorar la precisión de los resultados (ofreciendo aquellos más relevantes para usuarios anteriores)

Más recientemente ha estudiado el modo de extraer pseudo-folksonomías a partir de los conjuntos de consultas

Sistemas y servicios informáticos para Internet (2007/08)
Oviedo, 3, 4 y 5 de Marzo de 2008
Departamento de Informática
Web Semántica

Minería web *(Web Mining)*

A mediados de 2006 *Microsoft* financió una serie de proyectos de investigación sobre un conjunto de 15 millones de consultas

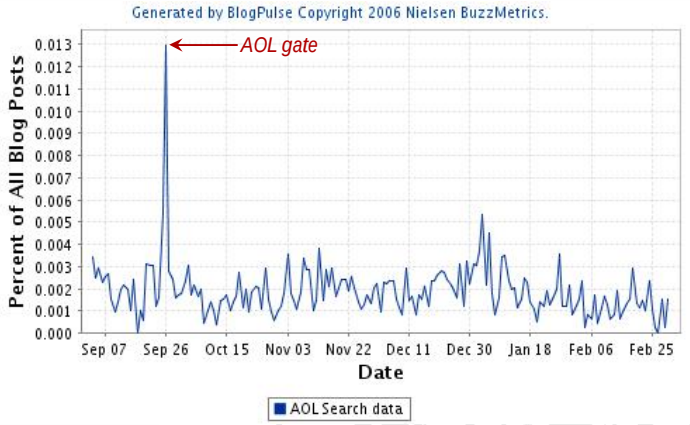
Poco después *AOL* liberó un archivo con datos sobre 20 millones de consultas correspondientes a 650.000 usuarios (miniescándalo)…

Rodaron cabezas, el sitio web fue eliminado en cuestión de horas y los datos…

…no tardaron en ser replicados, hoy sobreviven en algunos *mirrors* y redes *P2P*. Si os interesan quizás podáis descargarlos… AOL-**DATA.TGZ**

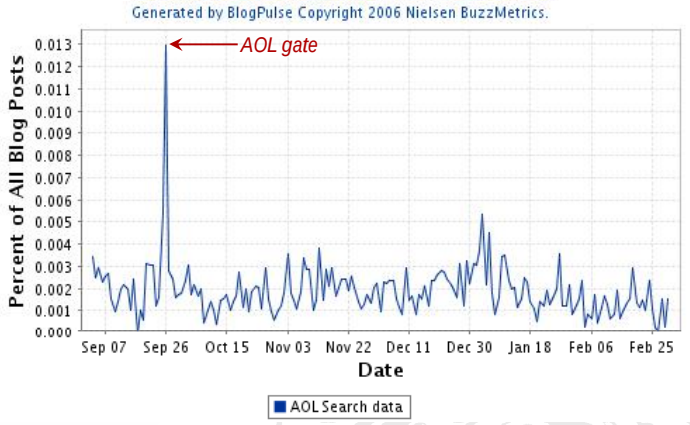
En 2007 *Microsoft* volvió a financiar proyectos de investigación relacionados con *semantic computing* e *internet economics* proporcionando, aparentemente, el mismo *log* de consultas que en 2006.

Análisis de tendencias

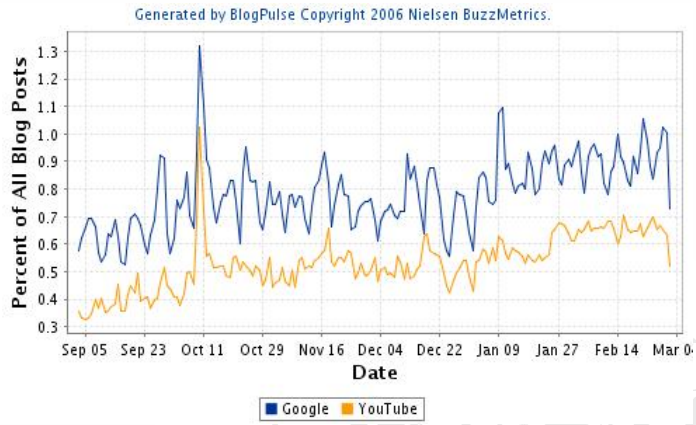


Sistemas y servicios informáticos para Internet (2007/08)
Oviedo, 3, 4 y 5 de Marzo de 2008
Departamento de Informática
Web Semántica

Análisis de tendencias

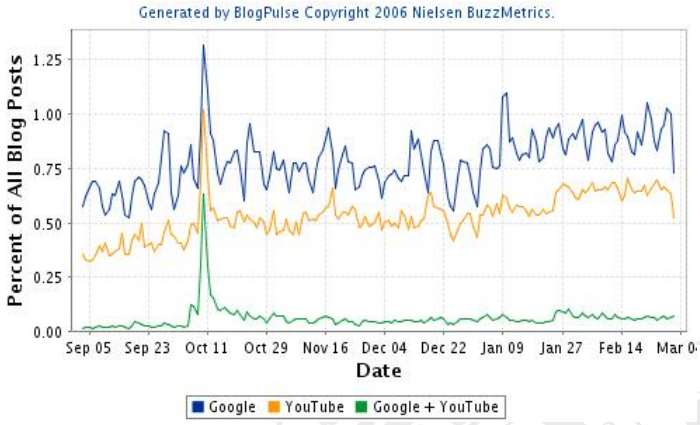


Análisis de tendencias



Sistemas y servicios informáticos para Internet (2007/08)
Oviedo, 3, 4 y 5 de Marzo de 2008
Departamento de Informática
Web Semántica

Análisis de tendencias



Análisis de tendencias

La frecuencia absoluta ya es un indicador interesante
La co-ocurrencia de términos es aún más interesante

Sistemas y servicios informáticos para Internet (2007/08)
Oviedo, 3, 4 y 5 de Marzo de 2008
Departamento de Informática
Web Semántica

Normalized *Google* Distance

Cilibrasi, R. y Vitanyi, P. 2005, *Automatic Meaning Discovery Using Google*, **http://arxiv.org/abs/cs.CL/0412098** ★

> *The rise of the world-wide-web has enticed millions of users to type in trillions of characters to create billions of web pages of on average low quality contents. The sheer mass of the information available about almost every conceivable topic makes it likely that extremes will cancel and the majority or average is meaningful in a low-quality approximate sense.*

Normalized Google Distance (NGD)

$$NGD(x,y) = \frac{\max\{\log f(x), \log f(y)\} - \log f(x,y)}{\log M - \min\{\log f(x), \log f(y)\}}$$

Normalized *Google* Distance

Distancias *NGD* entre algunos países

portugal	**0**	**0,02**	**0,21**	**0,07**	**0,23**	**0,09**	**0,11**
spain	**0,02**	**0**	**0,18**	**-0,01**	**0,15**	**0,10**	**0,12**
france	**0,21**	**0,18**	**0**	**0,16**	**-0,01**	**0,20**	**0,30**
italy	**0,07**	**-0,01**	**0,16**	**0**	**0,12**	**0,11**	**0,16**
germany	**0,23**	**0,15**	**-0,01**	**0,12**	**0**	**0,17**	**0,24**
belgium	**0,09**	**0,10**	**0,20**	**0,11**	**0,17**	**0**	**0,00**
netherl.	**0,11**	**0,12**	**0,30**	**0,16**	**0,24**	**0,00**	**0**

Primer problema: *Google* sólo proporciona estimaciones sobre el número total de documentos que contienen una palabra y, muchas veces, es una aproximación demasiado "gruesa" (p.ej. **spain** 311x10^6, **italy** 303x10^6, **spain italy** 330x10^6)

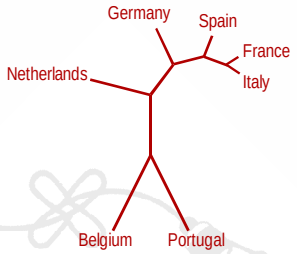
Sistemas y servicios informáticos para Internet (2007/08)
Oviedo, 3, 4 y 5 de Marzo de 2008
Departamento de Informática
Web Semántica

Normalized *Google* Distance

Segundo problema:

¿Sobre qué "eje" se mide la distancia?

En este ejemplo, ¿población? ¿superficie? ¿PIB?



En resumen, interesante, inspiradora… Aún se necesita más trabajo…

La Web como *corpus*

Un *corpus* es una colección de documentos que muestran el uso real de la lengua natural

Pueden ser monolingües o multilingües y estos, a su vez, paralelos o comparables

Los *corpora* multilingües son un recurso fundamental para la construcción de sistemas estadísticos de traducción automática

- Brown, P.F. *et al.* 1990, "A Statistical Approach to Machine Translation", *Computational Linguistics*, vol. 16, no. 2

Viktor aprende inglés comparando dos guías turísticas de Nueva York

http://video.google.com/videoplay?docid=6934089019347797736

Sistemas y servicios informáticos para Internet (2007/08)
Oviedo, 3, 4 y 5 de Marzo de 2008
Departamento de Informática
Web Semántica

La Web como *corpus*

La traducción automática es sólo uno de los campos que puede beneficiarse de la utilización de la Web como *corpus* o, mejor dicho, de *corpora* extraídos de la Web… Sin embargo, es uno de los más espectaculares

Trabajos interesantes:

- Jones, R. y Ghani, R. 2000, "Automatically Building a Corpus for a Minority Language from the Web", en *Proceedings of the Student Workshop of the 38th Annual Meeting of the Association for Computational Linguistics*, pp. 29–36
- Resnik, P. y Smith, N.A. 2003, "The Web as a parallel corpus", *Computational Linguistics*, vol. 29, no. 3, pp. 349-380

  Kilgarriff, A. y Grefenstette, G. 2003, "Introduction to the special issue on the web as corpus", *Computational Linguistics*, vol. 29, no. 3, pp. 333-347

A tener en cuenta: *WAC (Web as a Corpus Workshop)*, *CLEANEVAL*

Para saber más…

Manning, C.D. y Schutze, H. 1999, *Foundations of statistical natural language processing*, MIT Press ★★★

Chakrabarti, S. 2002, *Mining the Web: Discovering Knowledge from Hypertext Data*, Morgan Kaufmann ★★★

Manning, C.D., Raghavan, P. y Schütze, H. 2007, *Introduction to Information Retrieval*, Cambridge University Press (*online*) ★★★

Sistemas y servicios informáticos para Internet (2007/08)
Oviedo, 3, 4 y 5 de Marzo de 2008
Departamento de Informática
Web Semántica

Evolución

Aplicando las técnicas que hemos comentado ¿qué nos depara el futuro?

¿Cómo pueden evolucionar los buscadores?

Echemos a volar la imaginación *"a la Berners-Lee"*… **J**

Evolución
El presente

Lenssen, P. 2007, "Evolution of a search Engine", *Google Blogoscoped*, **http://blog.outer-court.com/archive/2007-02-02-n25.html**

Sistemas y servicios informáticos para Internet (2007/08)
Oviedo, 3, 4 y 5 de Marzo de 2008
Departamento de Informática
Web Semántica

Evolución
El futuro inmediato

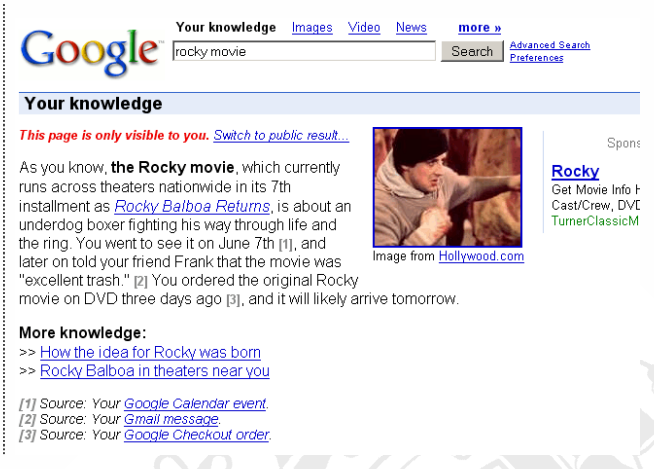
Evolución

El futuro inmediato

Sistemas y servicios informáticos para Internet (2007/08)
Oviedo, 3, 4 y 5 de Marzo de 2008
Departamento de Informática
Web Semántica

Evolución

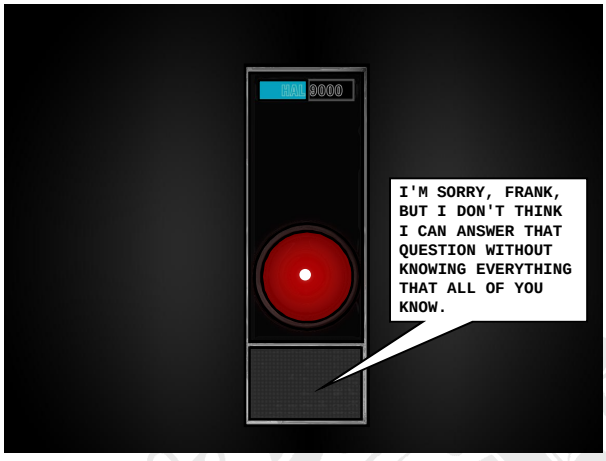
Verdadera personalización



Evolución
Inferencia

Sistemas y servicios informáticos para Internet (2007/08)
Oviedo, 3, 4 y 5 de Marzo de 2008
Departamento de Informática
Web Semántica

Evolución
... y más allá



STOP!

Por hoy estuvo bien…

¿Preguntas?

Lecturas de hoy

Broder, A. 2002, "A taxonomy of web search", *ACM SIGIR Forum*, vol. 36, no. 2, pp. 3-10.

Wu, J. y Aberer, K. 2003, "Swarm Intelligent Surfing in the Web", *ICWE 2003*, LNCS 2722, pp. 431–440

Baeza-Yates, R. et al. 2004, "Query recommendation using query logs in search engines", en *Current Trends in Database Technology*, LNCS 3268, p. 588-596.

Cilibrasi, R. y Vitanyi, P. 2005, *Automatic Meaning Discovery Using Google*, http://arxiv.org/abs/cs.CL/0412098

¿Quieres pasarte al lado oscuro?

Sistemas y servicios informáticos para Internet (2007/08)
Oviedo, 3, 4 y 5 de Marzo de 2008
Departamento de Informática
Web Semántica

I WANT YOU

for the DARK SIDE
ENLIST NOW

¿Qué cosas se han hecho hasta ahora o se están haciendo en el "lado oscuro"…?

Prototipos de sistemas de recuperación de información, resumen automático, identificación de idiomas.

Se he preparado un *corpus* paralelo inglés japonés alineado a nivel de sentencia que se siente muy solo…

Un compañero está explorando la forma de aplicar *swarm intelligence* al campo de la recuperación de información en la Web.

Otra compañera ha estudiado distintos algoritmos para obtener redes asociativas a partir de texto plano (un ejemplo a continuación).

Varios alumnos están trabajando en la forma de explotar la Web como un *corpus* libre de ruido sin necesidad de recurrir al *screen-scraping*.

I WANT YOU

Un *query log* es un archivo que contiene las consultas que los usuarios envían a un buscador.

avaliação carro construir instrumentos de musica misturas para aves tricot bonecos purificadores agua pensões porto	avaliação automovel fabricar instrumentos de musica misturas para periquitos coser malhas desenho animados filtros agua residenciais porto	carro and automovel are co-hyponym construir and fabricar are co-hyponyms periquito is hyponym of ave tricot can de defined as coser malha bonecos and desenhos aniamdos are co-hyponyms purificador and filtro are co-hyponyms pensão and residencial are co-hyponyms

Table 1: Example of extractable semantic links from sessions.

Seco, N. y Cardoso, N. 2006, "Detecting User Sessions in the Tumba! Query Log"

automáticas.

¿Relación con folksonomías y ontologías?

for the DARK SIDE
ENLIST NOW

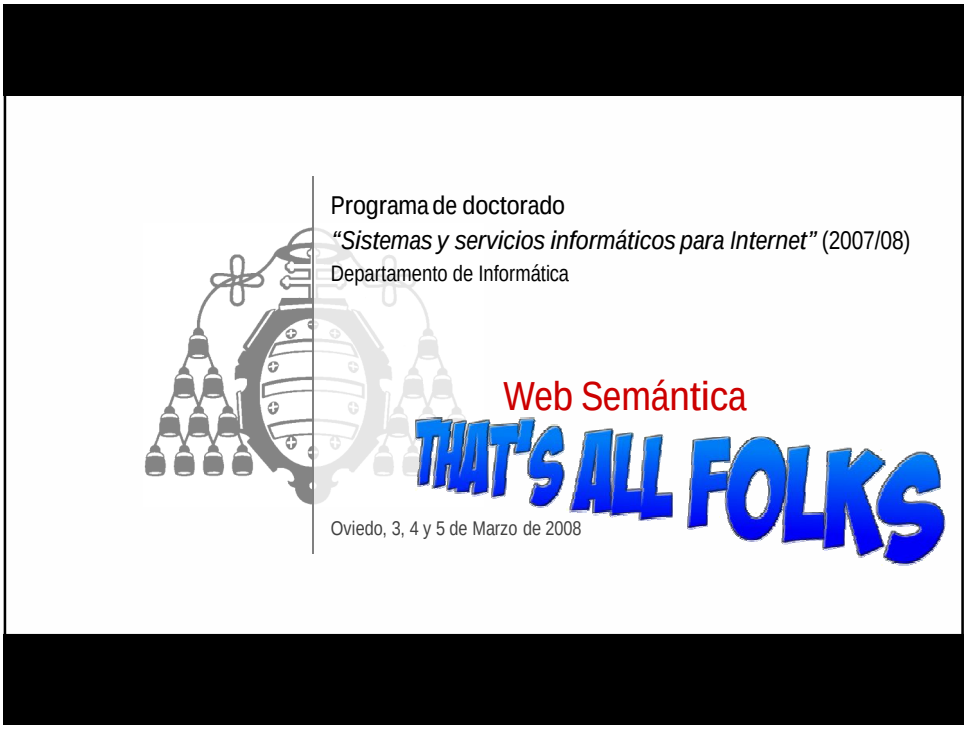
Sistemas y servicios informáticos para Internet (2007/08)
Oviedo, 3, 4 y 5 de Marzo de 2008
Departamento de Informática
Web Semántica

I WANT YOU

for the DARK SIDE
ENLIST NOW

Un ejemplo reciente, se generó una red asociativa para los artículos de la Wikipedia enlazados con el de Pablo Picasso.

¿Qué términos serán los que están más fuertemente asociados con **`Picasso`** de los 12.000 términos directamente relacionados?


Programa de doctorado
"Sistemas y servicios informáticos para Internet" (2007/08)
Departamento de Informática
Web Semántica
THAT'S ALL FOLKS
Oviedo, 3, 4 y 5 de Marzo de 2008