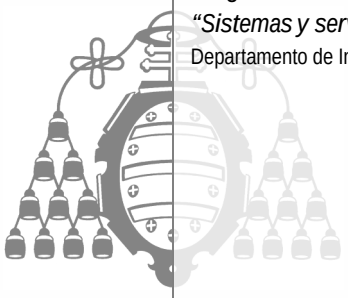


Programa de doctorado

"Sistemas y servicios informáticos para Internet" (2007/08)

Departamento de Informática

Web Semántica

Oviedo, 3, 4 y 5 de Marzo de 2008

Antes de empezar…

Evaluación del curso

- La evaluación del curso consistirá en la realización de un trabajo sobre algún tema relacionado con la Web Semántica y consistente en la escritura y presentación de una comunicación a un congreso simulado
- La comunicación (5 páginas) se presentará durante las clases del curso de doctorado y tiene que ser admitida por los profesores del curso
- La presentación será de 15 minutos con otros 15 minutos para preguntas
- Los alumnos que no puedan asistir a las clases y al congreso simulado presentarán un trabajo con formato de artículo de revista (LNCS, 15 páginas) a entregar el 21 de Abril de 2008

Más información en: `http://www.di.uniovi.es/~labra/cursos/Doc08UniOvi/`

Calendario

L 3, M 4 y X 5 de marzo (Dani Gayo)
J 6 y V 7 de marzo (sin clase)
L 10 de marzo (Labra)
M 11 y X 12 de marzo (videoconferencia en Gijón)
J 13 y V 14 de marzo (Labra)
X 26, J 27 y V 28 de marzo (Labra + presentación de trabajos)

¿Qué vamos a ver los próximos tres días?

La *Web-de-datos*

La Web como fuente de información

Presente y futuro de la Web

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

La *Web-de-datos*

Cuando éramos suficientemente jóvenes…

Advocatus diaboli

Web Semántica es esto…

¿Es esto Web Semántica?

No hay cuchara…

En suma…

Cuando éramos suficientemente jóvenes...

Suiza, 1989

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

Cuando éramos suficientemente jóvenes...

Berners-Lee, T. 1989, *Information Management: A Proposal*, Informe técnico, CERN. ★

> *Keywords can be nodes which stand for a concept. A keyword node is then no different from any other node. One can link documents, etc., to keywords. One can then find keywords by finding any node to which they are related. In this way, documents on similar topics are indirectly linked, through their key concepts. A keyword search then becomes a search starting from a small number of named nodes, and finding nodes which are close to all of them.*

Cuando éramos suficientemente jóvenes…

¡Genial! ¿Dónde hay que firmar?

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

Cuando éramos suficientemente jóvenes…

Massachusetts (EE.UU.), 12 años después…

Cuando éramos suficientemente jóvenes...

Berners-Lee, T. *et al.* 2001, "The Semantic Web", *Scientific American*, vol. 284, no. 5, pp. 34-43. ★

> *The Semantic Web will bring structure to the meaningful content of Web pages, creating an environment where **software agents** roaming from page to page can readily carry out **sophisticated tasks** for users.*
>
> ...
>
> *The Semantic Web is not a separate Web but an extension of the current one, in which information is given **well-defined meaning**, better enabling computers and people to work in cooperation.*
>
> ...
>
> *For the semantic web to function, computers must have access to structured collections of information and sets of **inference rules** that they can use to conduct **automated reasoning**.*

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

Cuando éramos suficientemente jóvenes...

Berners-Lee, T. *et al.* 2001, "The Semantic Web", *Scientific American*, vol. 284, no. 5, pp. 34-43.

> ...
>
> *The Semantic Web will enable machines to **COMPREHEND** semantic documents and data, not human speech and writings."*
>
> ...
>
> *[...] the third basic component of the Semantic Web, collections of information called **ontologies**.*

Cuando éramos suficientemente jóvenes...

Berners-Lee, T. *et al.* 2001, "The Semantic Web", *Scientific American*, vol. 284, no. 5, pp. 34-43.

> ...
> *The Semantic Web will enable machines to COMPREHEND semantic documents and data, not human speech and writings."*
> ...
> *[...] the third basic component of the Semantic Web, collections of information called ontologies.*

An ontology is a document or file that formally defines the relations among terms. The most typical kind of ontology for the Web has a taxonomy and a set of inference rules.

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

Cuando éramos suficientemente jóvenes...

¡Genial! ¿Dónde hay que firmar?

Cuando éramos suficientemente jóvenes…

Reino Unido, 5 años después…

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

Cuando éramos suficientemente jóvenes…

Shadbolt, N. *et al.* 2006, "The Semantic Web Revisited", *IEEE Intelligent Systems*, vol. 21, no. 3, pp. 96-101.

> *The Semantic Web is a Web of actionable information—information derived from data through a semantic theory for interpreting the symbols. The semantic theory provides an account of "meaning" in which the logical connection of terms establishes interoperability between systems. [...] This simple idea, however, remains largely unrealized.*

Cuando éramos suficientemente jóvenes…

Shadbolt, N. *et al.* 2006, "The Semantic Web Revisited", *IEEE Intelligent Systems*, vol.21, no.3, pp. 96-101.

> *The Scientific American article assumed that this would be straightforward, but it's still difficult to achieve in today's Web.*
>
> …
>
> *Because we haven't yet delivered large-scale, agent-based mediation, some commentators argue that the Semantic Web has failed to deliver. We argue that agents can only flourish when standards are well established [...]*
>
> …
>
> *The ontologies that will furnish the semantics for the Semantic Web must be developed, managed, and endorsed by practice communities.*

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

Cuando éramos suficientemente jóvenes…

Aquí y ahora…

Cuando éramos suficientemente jóvenes...

Resulta que no era pa' hoy...

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

Advocatus diaboli

Soergel, D. 1999, "The rise of ontologies or the reinvention of classification", *Journal of the American Society for Information Science*, vol.50, no.12, pp. 1119-1120.

> *Ontologies are developed in many communities of research and practice. Unfortunately, there is little communication and mutual learning; thus, efforts are fragmented, resulting in considerable reinvention and less than optimal products.*

Bates, M.J. 2002, "After the Dot-Bomb: Getting Web Information Retrieval Right This Time", *First Monday*, vol. 7, no. 7 ★

> *Succumbing to the "ontology" fallacy...*

Shirky, C. 2005. "Ontology is Overrated: Categories, Links and Tags", <http://www.shirky.com/writings/ontology_overrated.html>

Advocatus diaboli

Abelson, H. 2005

> *[...] A lot of the enthusiasm around the SemWeb reminds me of the AI hullabaloo of the 1980s. [...] Over the past 20 years, AI researchers have come to appreciate the limitations of traditional knowledge representation techniques. It seems that statistical methods and machine learning have proven more productive than reasoning based on ontologies. [...]*

Hendler, J. 2006, *"The Dark Side of the Semantic Web"*

> *[...] the Semantic Web vision of Tim's, before Ora and I polluted it with all this ontology stuff [...]*

Antoniou, G. 2007, charla invitada durante *MTSR'07*

> *The semantic web may fail but semantic web technologies will stay.*

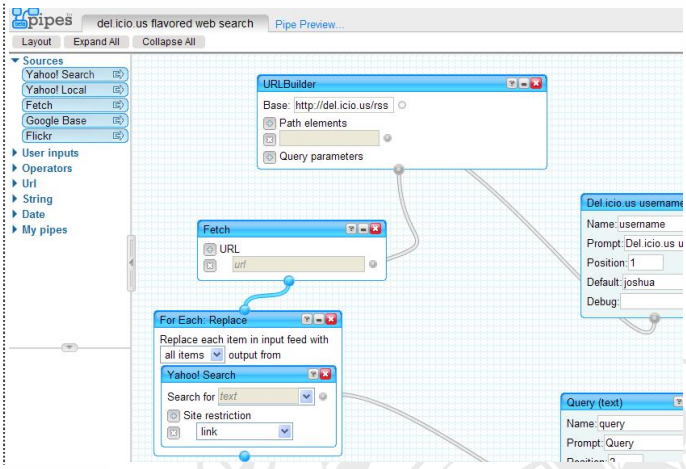
Sistemas y servicios informáticos para Internet (2007/08)
Oviedo, 3, 4 y 5 de Marzo de 2008
Departamento de Informática
Web Semántica

Web Semántica es esto…

Según el W3C:

> *The Semantic Web is a web of data.*
>
> *…*
>
> *The Semantic Web is about two things. It is about common formats for integration and combination of data drawn from diverse sources, where on the original Web mainly concentrated on the interchange of documents. It is also about language for recording how the data relates to real world objects. That allows a person, or a machine, to start off in one database, and then move through an unending set of databases which are connected not by wires but by being about the same thing.*

¿Es esto Web Semántica?



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

¿Es esto Web Semántica?

Yahoo! pipes

> *Pipes is an interactive feed aggregator and manipulator. Using Pipes, you can create feeds that are more powerful, useful and relevant.*

Tim O'Reilly (febrero 2007)

> *Yahoo!'s new Pipes service is a milestone in the history of the internet. It's a service that generalizes the idea of the mashup, [...] [it] allows you to connect internet data sources, process them, and redirect the output.*

¿Es esto Web Semántica?

¿Hay combinación de datos?

Sí

¿Hay formatos comunes?

Sí (*RSS* y *JSON*)

¿Puede una persona/máquina moverse de una "base de datos" a otra sin solución de continuidad?

Más o menos…

¿Usa alguna de las tecnologías propuestas por el W3C?

Bueeeno… RSS 1.0 se construye a partir de RDF que es una tecnología W3C.

¿Es Web Semántica?

No en el sentido ortodoxo, pero tal vez sea un paso para la *Web-como-base-de-datos*

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

No hay cuchara…

Olvidemos el nombre, Web Semántica…

Olvidemos la ortodoxia (ontologías, RDF, etc.)

¿Qué perseguimos?

La *Web-como-base-de-datos*

¿Alguien más, aparte de la *gente-de-la-Web-Semántica*, busca más o menos lo mismo?

Claro **J**

¿Por ejemplo? Los siguientes campos en solitario o combinados:

- Extracción de información, de entidades, de términos…
- Respuesta de preguntas.
- Recuperación de información.
- Resumen automático.
- Traducción automática.
- …

No hay cuchara…

Extracción de información *(Information Extraction)*

El objetivo fundamental de la extracción de información es obtener información estructurada (fundamentalmente entidades y relaciones entre las mismas) a partir de texto poco o nada estructurado.

Extracción de entidades *(Entity Extraction, Named-Entity Recognition)*

Una subtarea dentro del campo de extracción de información cuyo objetivo es localizar en un texto libre aquellos fragmentos que se corresponden con nombres de personas, organizaciones, lugares, etc.

Extracción de términos *(Term Extraction)*

Otra subtarea del campo de extracción de información. Su objetivo es localizar términos (palabras o frases) relevantes para el tema de una colección de documentos.

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

No hay cuchara…

Respuesta a preguntas *(Question Answering, QA)*

Un sistema de respuesta a preguntas es aquel que permite a los usuarios plantear una pregunta en lenguaje natural y recibir una respuesta concisa (no un documento) con suficiente contexto como para verificar su validez.

`http://start.csail.mit.edu/`

Recuperación de información *(Information Retrieval, IR)*

El término recuperación de información hace referencia, en general, al estudio de sistemas automáticos que permitan a un usuario determinar la existencia o inexistencia de documentos (esto es, textos) relativos a una necesidad de información formulada habitualmente como una consulta.

Resumen automático *(Automatic Summarization)*

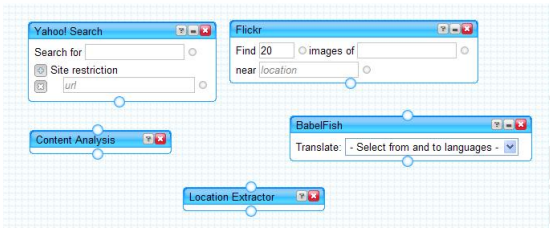
Las técnicas de resumen automático tienen como misión obtener a partir de un documento o conjunto de documentos un único texto mucho más corto que aún contenga los aspectos más relevantes de los originales.

Traducción automática *(Machine Translation)*

El objetivo de la traducción automática es bastante obvio: traducir, sin intervención humana, un texto de un idioma a otro. En la actualidad el paradigma más empleado es el estadístico (empleando modelos generados a partir de grandes cantidades de texto bilingüe).

No hay cuchara…

Muchas de estas tareas son módulos en *Yahoo! pipes…*



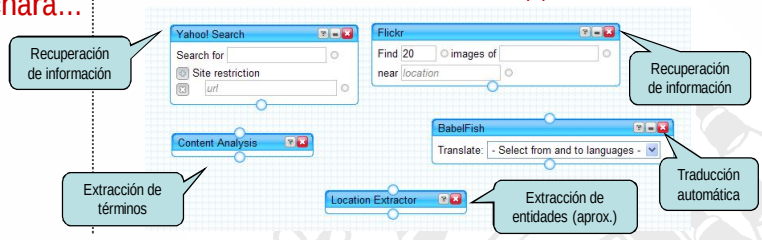
Sistemas y servicios informáticos para Internet (2007/08)
Oviedo, 3, 4 y 5 de Marzo de 2008
Departamento de Informática
Web Semántica

Traducción automática *(Machine Translation)*

El objetivo de la traducción automática es bastante obvio: traducir, sin intervención humana, un texto de un idioma a otro. En la actualidad el paradigma más empleado es el estadístico (empleando modelos generados a partir de grandes cantidades de texto bilingüe).

No hay cuchara…

Muchas de estas tareas son módulos en *Yahoo! pipes…*



En suma…

Los próximos 3 días vamos a hablar de todas estas técnicas que pueden conducirnos a esa *Web-de-datos* además de otras varias para extraer conocimiento de la Web.

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

En suma…

clustering folksonomía *click-through data* modelo booleano
modelo vectorial etiquetado
recuperación de información *PageRank*
evaluación *stemming* *relevance feedback*
relevancia *pseudo-relevance feedback*
búsquedas en la Web *HITS* *NGD* *tf*idf*

La Web como fuente de información

- Pero, ¿cuál es el problema real?
- De aquellos polvos…
- …vienen estos lodos
- Encontrar información en la Web (antes de *Google*)
- Recuperación de información en dos palabras (o más…)
- Hitos en recuperación de información (hasta *Google*)
- ¿Por qué las técnicas *IR* clásicas no funcionan bien en la Web?
- La Web es un grafo
- *PageRank*
- Búsquedas en la Web con *PageRank*
- ¿Son adecuados los buscadores modernos?
- (Más) Problemas del *ranking* basado en hiperenlaces
- No hay talla única…

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

Pero, ¿cuál es el problema real?

Algunas cifras (la mayoría obsoletas):

- Desde 1981 se han generado más de 845 x 10^6 de mensajes en *USENET*
- *Reuters* produce 11 x 10^3 artículos diarios
- *Springer* publicó en 2003 90 x 10^6 palabras en textos científicos
- El tamaño real de la Web es desconocido:
  - La Web superficial tiene más de 4 x 10^9 documentos
  - La Web oculta puede ser entre 2x y 500x
- Existen más de 70 x 10^6 *blogs*
- *flickr* contiene más de 17 x 10^6 fotografías y sus usuarios añaden cada día 1,2 x 10^6 etiquetas
- …

Pero, ¿cuál es el problema real?

Alvin Tofler (1970) definió la "sobrecarga de información" como la condición que se deriva de la incapacidad de la mente humana para enfrentarse a demasiados elementos de información simultáneamente.

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

Pero, ¿cuál es el problema real?

Se puede decir más alto pero no más claro…

> *Me gusta Internet. De verdad, me encanta. Siempre que necesito algo de shareware o ver qué tiempo hace en Bogotá soy el primero en hacer zumbar el módem. Pero como fuente de información, es una mierda. Encuentras millones de datos peleándose por hacerse oír, ver y descargar y parece que todo lo que quieres saber está aplastado por la multitud.*

Michael Marshall (2002), *The Straw Men* (traducción libre)

Pero, ¿cuál es el problema real?

(Intermedio)

Véronis, J. 2007, "Search: Google-Yahoo Comparison"
<http://aixtal.blogspot.com/2007/11/search-google-yahoo-comparison.html>

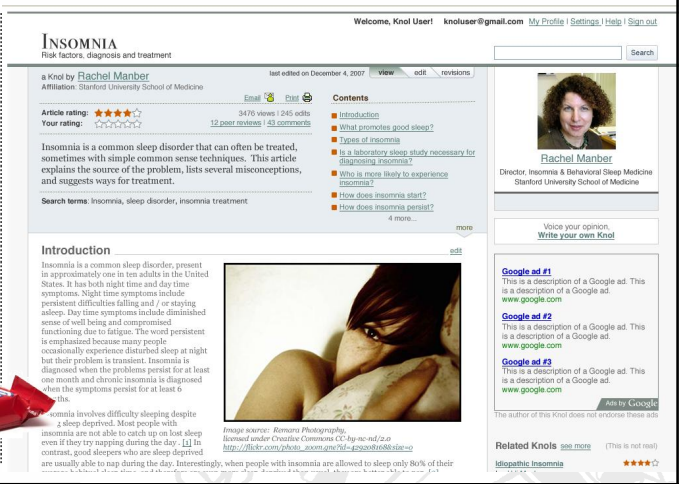
> The most surprising result came from the use of Wikipedia. This use was marginal in December 2005. At the time, for all 10 results on the first page, 2% of the links proposed by Google and 4% of those proposed by Yahoo came from Wikipedia.
>
> The strategies have changed completely. Today 27% of Google's results on the first link alone come from Wikipedia, as do 31 % of Yahoo's.

Reflexionad sobre esto…

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

Pero, ¿cuál es el problema real?

(Intermedio)



Pero, ¿cuál es el problema real?

(Intermedio)

Introduction

Insomnia is a common sleep disorder, present in approximately one in ten adults in the United States. It has both night time and day time symptoms. Night time symptoms include persistent difficulties falling and / or staying asleep. Day time symptoms include diminished sense of well being and compromised functioning due to fatigue. The word persistent is emphasized because many people occasionally experience disturbed sleep at night but their problem is transient. Insomnia is diagnosed when the problems persist for at least one month and chronic insomnia is diagnosed when the symptoms persist for at least 6 months.

Insomnia involves difficulty sleeping despite being sleep deprived. Most people with [illegible]somnia are not able to catch up on lost sleep if they try napping during the day . [1] In [illegible]st, good sleepers who are sleep deprived

This is an old revision of this page, as edited by Splash (Talk | contribs) at 17:48, 9 December 2007. It may differ significantly from the current revision.

(diff) ← Older revision | current version (diff) | Newer revision → (diff)

This article may require cleanup to meet Wikipedia's quality standards.
Please improve this article if you can. (June 2007)

This article is about the sleeping disorder. For other uses, see Insomnia (disambiguation).

Insomnia is a sleeping disorder characterized by the inability to fall asleep and/or the inability to remain asleep for a reasonable amount of time. Insomniacs have been known to complain about being unable to close their eyes or "rest their mind" for more than a few minutes at a time. Both organic and non-organic insomnia constitute a sleep disorder.[1][2]

Insomnia	
Classification & external resources	
ICD-10	F51.0, G47.0
ICD-9	307.42, 307.41, 780.51, 780.52
DiseasesDB	26877
eMedicine	med/2698
MeSH	D007319

According to the U.S. Department of Health and Human Services, approximately 60 million Americans suffer from insomnia each year.[3] Insomnia tends to increase with age and affects about 40 percent of women and 30 percent of men.[4]

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

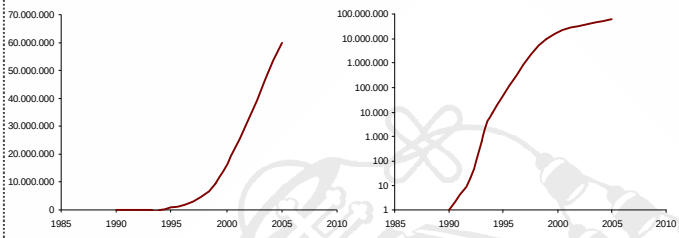
De aquellos polvos...

- Propuesta original para la Web (Berners-Lee, 1989)
  - Evitar pérdida de información
  - Facilitar acceso a toda la información
- Características que facilitaron crecimiento de la Web
  - Naturaleza distribuida (documentos pueden residir en distintas máquinas)
  - Hiperenlaces
  - Sistema tanto más útil cuantos más documentos contenga
- Reflexiones...
  - Búsqueda por palabras clave es un problema
  - En la propuesta original los conceptos son nodos idénticos a los documentos
- Desarrollo inicial de la Web
  - No hay nodos conceptuales, sólo documentos
  - No se implementa método alguno para buscar información

… vienen estos lodos

Encuentras millones de datos peleándose por hacerse ver y parece que todo lo que quieres saber está aplastado por la multitud.

Evolución en el número de servidores web…



Encontrar información en la Web (antes de *Google*)

Directorios

Bases de datos de enlaces organizados en categorías. Los enlaces suelen ser enviados por los responsables del sitio web y pueden existir editores que organicen la información disponible.

Por ejemplo, *CERN* (extinto) *NCSA* (extinto), *Yahoo!*, *ODP/Dmoz*

Problemas

- Muchos sitios web no notifican a los índices de su existencia
- No consiguen indexar la Web al ritmo que crece
- Recuperación de información "tradicional"
- Superabundancia de resultados y escasa relevancia

Encontrar información en la Web (antes de *Google*)

Buscadores

Artefactos *software* que exploran la Web almacenando en una base de datos parte o todo el texto de los documentos que analizan. Al ir procesando documentos se crea un índice que emplea las palabras que aparecen en cada página web. Cuando un buscador recibe una consulta toma las palabras utilizadas por el usuario y obtiene los documentos indexados por las mismas.

Por ejemplo, *ALIWEB*, *WebCrawler*, *Lycos* (extintos, permanecen las marcas)

Problemas

- Cobertura: la base de datos de cada buscador apenas representaba 1/3 de la Web
- Recuperación de información "tradicional"
- Superabundancia de resultados y escasa relevancia

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

Encontrar información en la Web (antes de *Google*)

Google cambió el panorama gracias al algoritmo *PageRank*

Para saber qué cambió, antes hay que entender cómo funciona un sistema de recuperación de información "tradicional"

Recuperación de información en dos palabras (o más…)

El término "recuperación de información" (*information retrieval* o *IR*) hace referencia al conjunto de procesos necesarios para representar, almacenar, buscar y encontrar información relevante para las consultas de los usuarios.

Un sistema de recuperación de información no informa al usuario, simplemente le indica la existencia (o inexistencia) de documentos relativos a la consulta.

Aunque, en principio, *IR* podría referirse a diversas manifestaciones de la información como imágenes, audio, texto, etc. se acepta generalmente que la "recuperación de información" se ocupa únicamente de información textual.

"La recuperación de información es un proceso de ensayo y error… Una consulta no es más que una suposición acerca de los atributos que se espera tenga el documento deseado. En general, se emplea la respuesta del sistema para corregir esa suposición inicial en posteriores intentos." (Swanson 1977)

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

Hitos en recuperación de información (hasta *Google*)

1950s

★ Primera descripción de un sistema *IR* automático. Utilización de la frecuencia de aparición de un término para determinar su relevancia, uso de *stoplists*. Luhn, H.P. 1957, "A Statistical Approach to Mechanized Encoding and Searching Information", *IBM Journal of Research and Development*, vol. 1, no. 4, pp. 309-317.

★ Primera propuesta para un sistema de resumen automático. Luhn, H.P. 1958, "The Automatic Creation of Literature Abstracts", *IBM Journal of Research and Development*, vol. 2, no. 2, pp. 159-165.

1960s

Primera alternativa "aritmética" a la búsqueda booleana. Maron, M.E. y Kuhns, K.L. 1960, "On relevance, probabilistic indexing and information retrieval", *Journal of the ACM*, vol. 7, no. 3, pp. 216-244.

Primer esfuerzo para la evaluación experimental de sistemas *IR*. Cleverdon, C.W. 1962, *Report on the Testing and Analysis of an Investigation into the Comparative Efficiency of Indexing Systems*, College of Aeronautics, Reino Unido.

Hitos en recuperación de información (hasta *Google*)

1960s

★ Se propone el modelo vectorial de documentos y medida coseno de similitud. Salton, G. y Lesk, M.E. 1965, "The SMART Automatic Document Retrieval System – An Illustration", *Communications of the ACM*, vol. 8, no. 6, pp. 391-398.

1970s

Se propone la *cluster hypothesis*, documentos estrechamente asociados tienden a ser relevantes para las mismas peticiones. Jardine, N. y van Rijsbergen, C.J. 1971, "The use of hierarchic clustering in information retrieval", *Information Storage and Retrieval*, vol. 7, pp. 217-240.

★ Introducción del concepto *idf (inverse document frequency)*. Spärck-Jones, K. 1972, "A statistical interpretation of term specificity and its application in retrieval", *Journal of Documentation*, vol. 28, no. 1, pp. 11-21.

Se propone el modelo probabilista de *IR*. Robertson, S.E. y Spärck-Jones, K. 1976, "Relevance weighting of search terms", *Journal of the ASIS*, vol. 27, no. 3, pp. 129-146.

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

Hitos en recuperación de información (hasta *Google*)

1970s

Por primera vez se señala la naturaleza interactiva de los sistemas *IR*. Swanson, D.R. 1977, "Information retrieval as a trial-and-error process", *Library Quarterly*, vol. 47, no. 2.

Primera colección moderadamente grande, *NPL* (11.500 documentos). Spärck-Jones, K. y Webster, C.A. 1979, *Research in Relevance Weighting*, Informe técnico, University of Cambridge.

1980s

★ Se inventa el primer algoritmo de *stemming*. Porter, M.F. 1980, "An algorithm for suffix stripping", *Program*, vol. 14, no. 3, pp. 130-137.

Se inventan los mapas auto-organizados. Kohonen, T. 1982, "Self-organized formation of topologically correct feature maps", *Biological Cybernetics*, 43, pp. 59-69.

Probabilidad de coincidencia entre dos individuos en el uso de la misma palabra para identificar un concepto está entre el 10 y el 20%. Furnas, G.W., Landauer, T.K., Gómez, L.M. y Dumais, S.T. 1987, "The vocabulary problem in human system communication", *Communications of the ACM*, vol. 30, no. 11, pp. 964-971.

Hitos en recuperación de información (hasta *Google*)

1980s

- Se inventa la Semántica Latente. Dumais, S.T., Furnas, G.W., Landauer, T.K., Deerwester, S. y Harshman, R. 1988, "Using Latent Semantic Analysis to improve access to textual information", en *Human Factors in Computing Systems*, CHI'88 Conference Proceedings, pp. 281-285.
- Se inventa la Web. Berners-Lee, T. 1989, *Information Management: A Proposal*, Informe técnico, CERN.

1990s

Se inventan las *Support Vector Machines*. Boser, B., Guyon, I. y Vapnik, V. 1992, "A training algorithm for optimal margin classifiers", en *Fifth Annual Workshop on Computational Learning Theory*, pp. 144-152.

- Se propone un método para detección de terminología. Dunning, T. 1993, "Accurate methods for the statistics of surprise and coincidence", en *Computational Linguistics*, vol. 19, no. 1, pp. 61-74.

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

Hitos en recuperación de información (hasta *Google*)

1990s

Se desarrollan los primeros buscadores web...
Koster, M. 1994, "ALIWEB – Archie-Like Indexing in the WEB", *Computer Networks and ISDN Systems*, vol. 27, no. 2, pp. 175-182.
Pinkerton, B. 1994, *"Finding what people want: Experiences with the WebCrawler"*
Mauldin, M.L. y Leavitt, J.R.R. 1994, *"Web Agent Related Research at the Center for Machine Translation"*

...Y los primeros índices
Filo, D. y Yang, J. 1994, *Yahoo!*

- Desarrollo de sistemas *IR* "tolerantes" por medio de n-gramas. Cavnar, W.B. 1994, "Using an n-gram-based document representation with a vector processing retrieval model", en *Proceedings of TREC-3*, pp. 269-277.

Primeros sistemas con *pseudo-relevance feedback*.
Robertson, S.E., Walker, S., Jones, S., Hancock-Beaulieu, M. y Gatford, M. 1994, "Okapi at TREC-2", en *Text REtrieval Conference*, pp. 21-34.
Buckley, C., Salton, G., Allan, J. y Singhal, A. 1995, "Automatic Query Expansion Using SMART: TREC-3", en *Text REtrieval Conference*, pp. 69-80.

Hitos en recuperación de información (hasta *Google*)

1990s

★ Se desarrolla la técnica *TextTiling* para detección de pasajes. Hearst, M.A. 1994, "Multi-Paragraph Segmentation of Expository Text", en *Proceedings of the 32nd Meeting of the Association for Computational Linguistics*, pp. 9-16.

Primeros pasos hacia la Web Semántica. Luke, S., Spector, L. y Rager, D. 1996, "Ontology-Based Knowledge Discovery on the World-Wide Web", en *Working Notes of the Workshop on Internet-Based Information Systems at the 13th National Conference on Artificial Intelligence (AAAI96)*.

1998 ANNO MACHINÆ INVENTÆ

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

Por hoy estuvo bien...

¿Preguntas?

Para mañana...

STOP!

Berners-Lee, T. 1989, *Information Management: A Proposal*, Informe técnico, CERN.

Koster, M. 1994, "ALIWEB – Archie-Like Indexing in the WEB", *Computer Networks and ISDN Systems*, vol. 27, no. 2, pp. 175-182.

Pinkerton, B. 1994, "Finding what people want: Experiences with the WebCrawler", [Online], Internet Archive, en *Electronic Proceedings of the "Second World Wide Web Conference '94: Mosaic and the Web"*, NCSA, Disponible en: <http://web.archive.org/web/20010904075500/http://archive.ncsa.uiuc.edu/SDG/IT94/Proceedings/Searching/pinkerton/WebCrawler.html>

Mauldin, M.L. y Leavitt, J.R.R. 1994, "Web Agent Related Research at the Center for Machine Translation", [Online], en *Proceedings of the ACM Special Interest Group on Networked Information Discovery and Retrieval*, Disponible en: <http://web.archive.org/web/19970607125802/http://fuzine.mt.cs.cmu.edu/mlm/signidr94.html>

¿En qué se diferencian las búsquedas en la Web de otro tipo de búsquedas?