Recuperación de Información

Actualidad y Retos

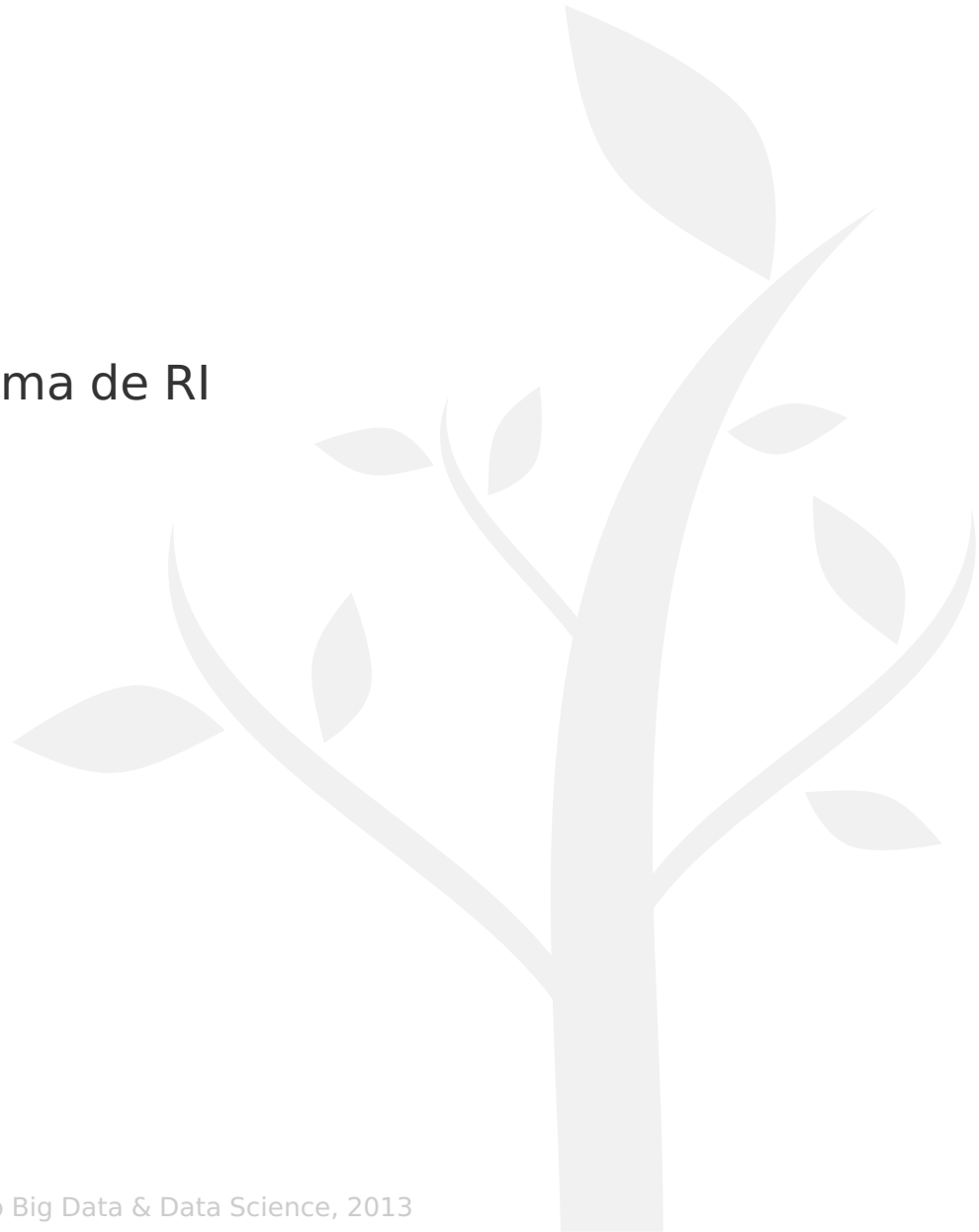
Dr. David E. Losada

Centro Singular de Investigación en Tecnoloxías da Información

Universidade de Santiago de Compostela

Curso de Verano Big Data & Data Science, 2013

Contenidos



Contenidos

Recuperación de Información (RI)

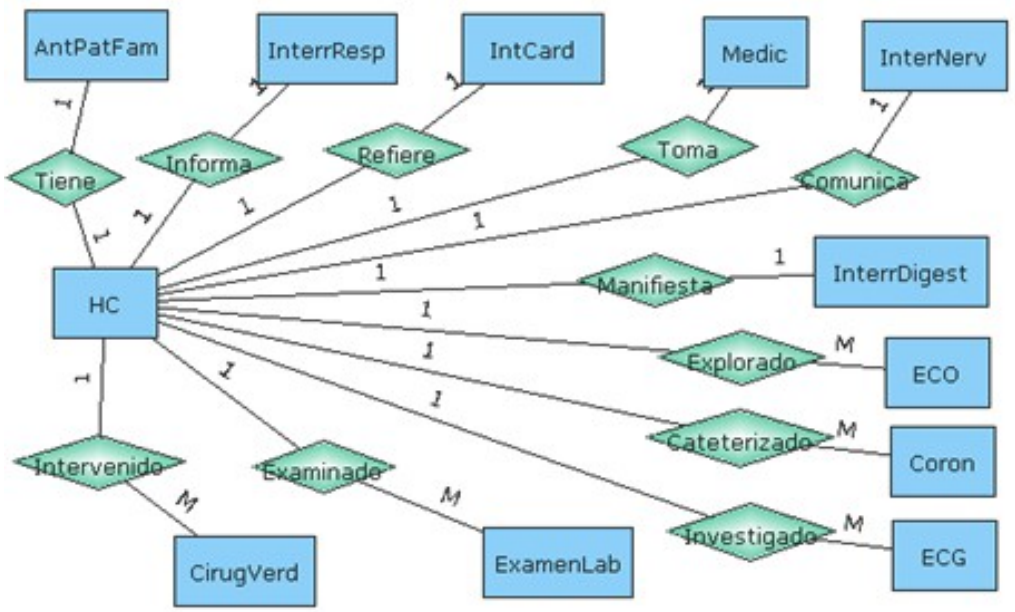
RI: Definición

RI es encontrar material (usualmente docs) de naturaleza no estructurada (usualmente texto) que satisfaga una necesidad de información en grandes colecciones almacenadas en computadoras

RI vs RD

Recuperación de Datos (RD)

- Estructurada (BD)
- Indexación por campos



RI vs RD

Recuperación de Datos (RD)

- Lenguaje Consulta Cerrado
- No ranking

	ID	FULLNAME	CCID	AGE	GENDER	BIRTHDAY	REGISTATIONDATETIME	ISDELETED
1	1	AAMIR HASAN	101	23	MALE	1990-01-01 00:00:00	2010-06-27 00:20:43.020	0
2	2	AMIR ALI	102	23	MALE	1993-02-01 00:00:00	2010-06-27 00:20:43.020	0
3	3	AHMED ALI	103	23	FEMALE	1994-08-01 00:00:00	2010-06-27 00:20:43.020	0
4	4	SONIA KHAN	104	23	FEMALE	1991-07-01 00:00:00	2010-06-27 00:20:43.020	0
5	5	AWAIS AHMED	105	23	MALE	1992-01-01 00:00:00	2010-06-27 00:20:43.023	0
6	6	AAMIR KHAN	106	23	MALE	1997-01-05 00:00:00	2010-06-27 00:20:43.023	0
7	7	SOBIA HINA	107	23	FEMALE	1988-01-01 00:00:00	2010-06-27 00:20:43.023	0
8	8	ADNAN KHAN	106	23	MALE	1987-01-01 00:00:00	2010-06-27 00:20:43.023	0
9	9	AAMIR HASAN	108	23	MALE	1997-04-01 00:00:00	2010-06-27 00:20:43.027	0
10	10	AAMIR HASAN	101	23	MALE	1990-01-01 00:00:00	2010-06-27 00:20:43.027	0
11	11	AAMIR KHAN	107	23	MALE	1990-01-01 00:00:00	2010-06-27 00:20:43.027	0

RI vs RD

Recuperación de Información (RI)

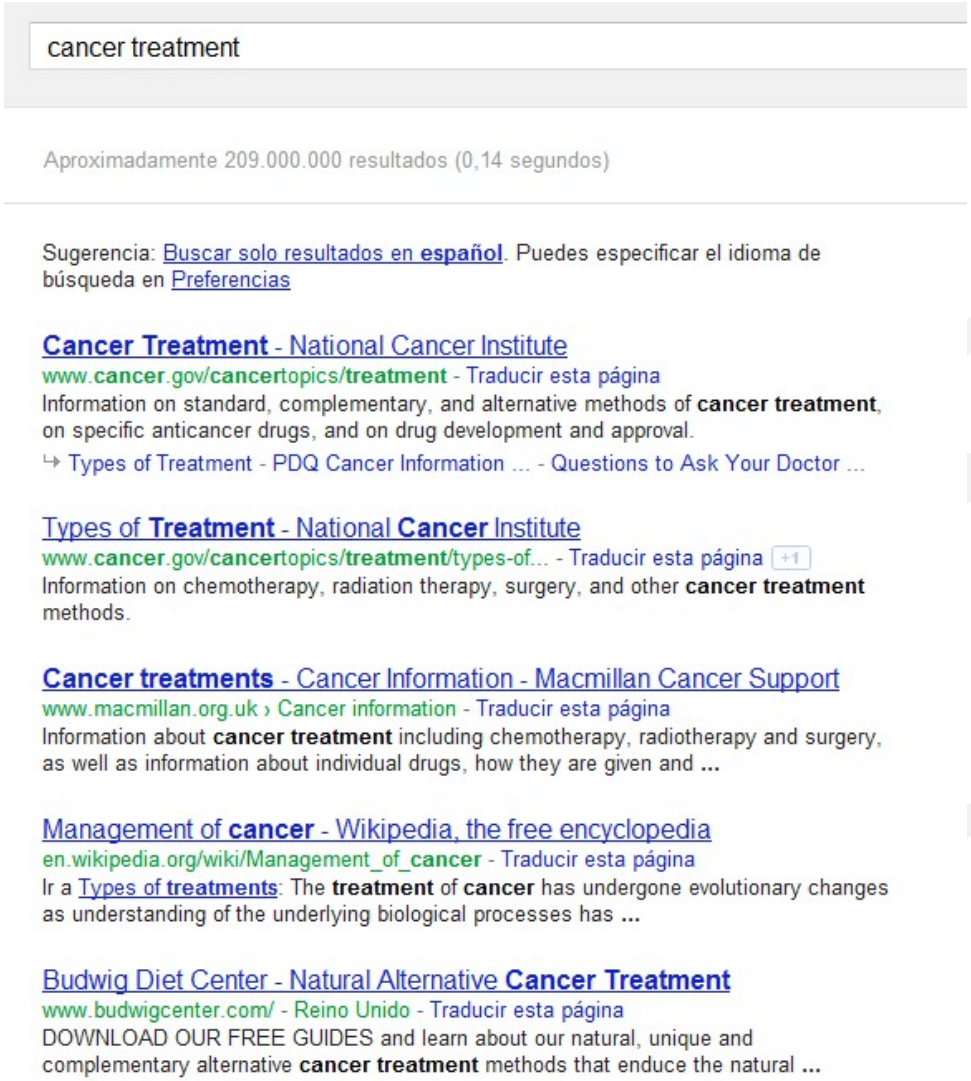
- No estructurada (p.e. texto)
- o Semi-estructurada (p.e. XML)
- Indexado Full-text


```
<?xml version="1.0" encoding="ISO-8859-1"?>
<report>
    <author>Ducyk, Philippe</author>
    <text>
        Specimen.
        The report body is inserted here.
        Specimen.
    </text>
    <patient>
        <id>123456789</id>
        <name>Patient, Fake A.</name>
        <birthdate>19501225</birthdate>
        <gender>M</gender>
        <zip>32610</zip>
        <city>Gainesville, FL</city>
        <country>USA</country>
    </patient>
    <request>
        <date>20070710</date>
        <info>Post operation check</info>
        <physician>Strangelove, Doctor</physician>
        <department>Orthopedics</department>
        <ward>PACU</ward>
        <state>H</state>
    </request>
    <exam>
        <code>1531</code>
        <description>Knee, left</description>
        <room>MOB9</room>
        <technician>DWI</technician>
        <radiologist>Duyck, Philippe</radiologist>
    </exam>
    <exam>
        <code>1521</code>
        <description>Femur, left</description>
        <room>MOB9</room>
        <technician>DWI</technician>
        <radiologist>Duyck, Philippe</radiologist>
    </exam>
</report>
```

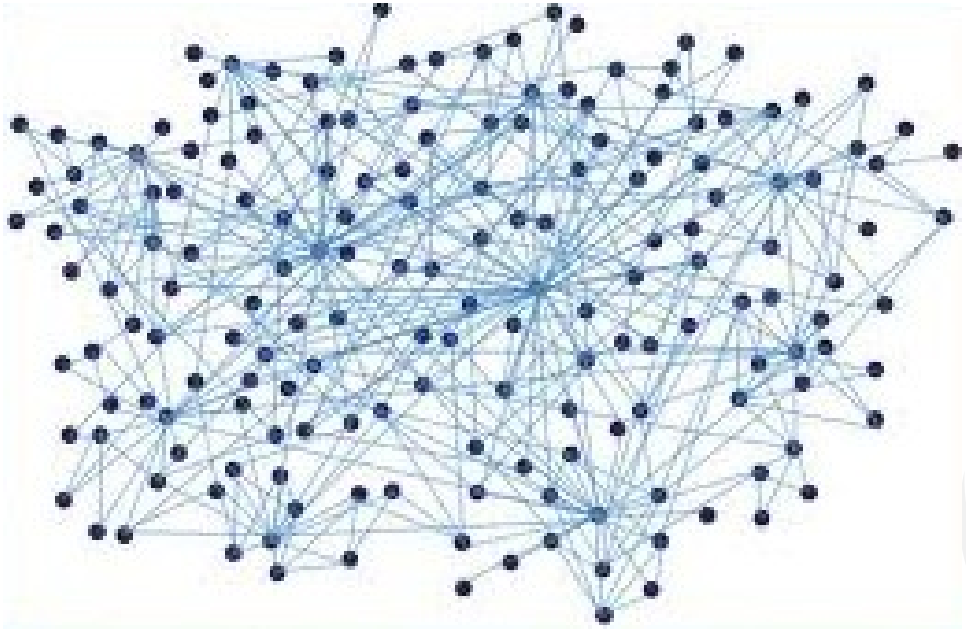
RI vs RD

Recuperación de Información (RI)

- Consultas libres
- Ranking por relevancia



Ejemplo paradigmático: La Web



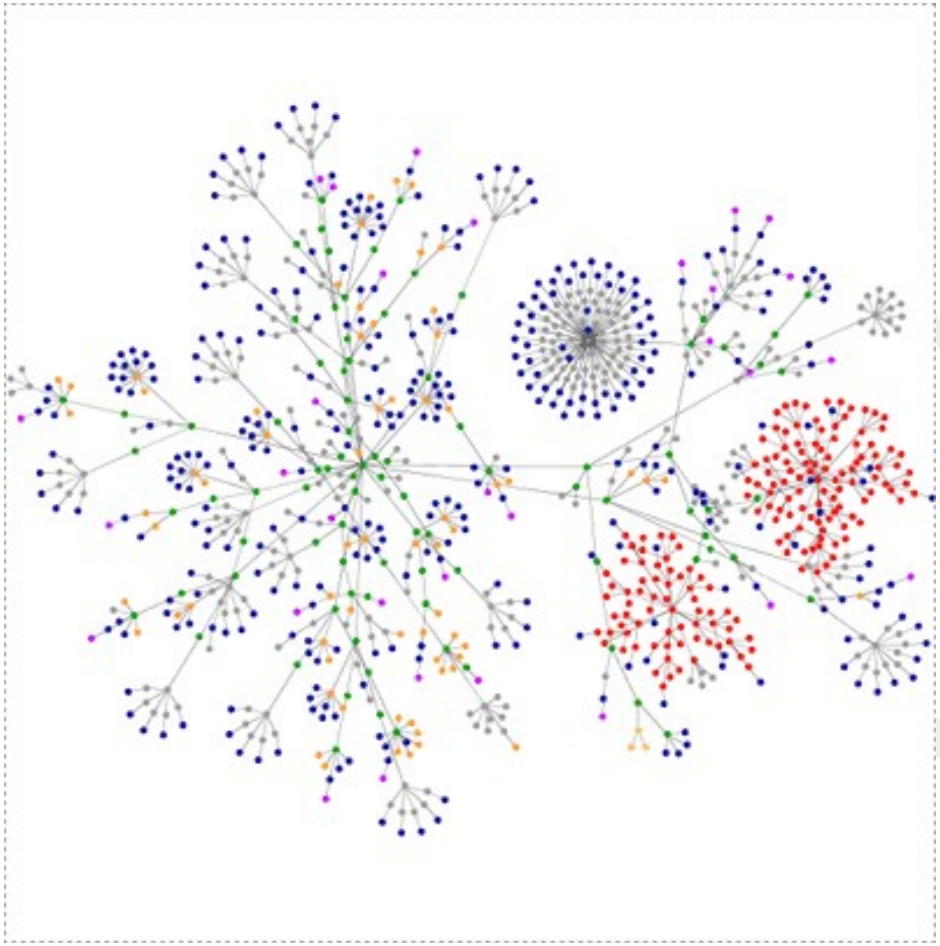
- Motores búsqueda tipo Google
- Crawling, Indexación, Recuperación

Ejemplo paradigmático: La Web

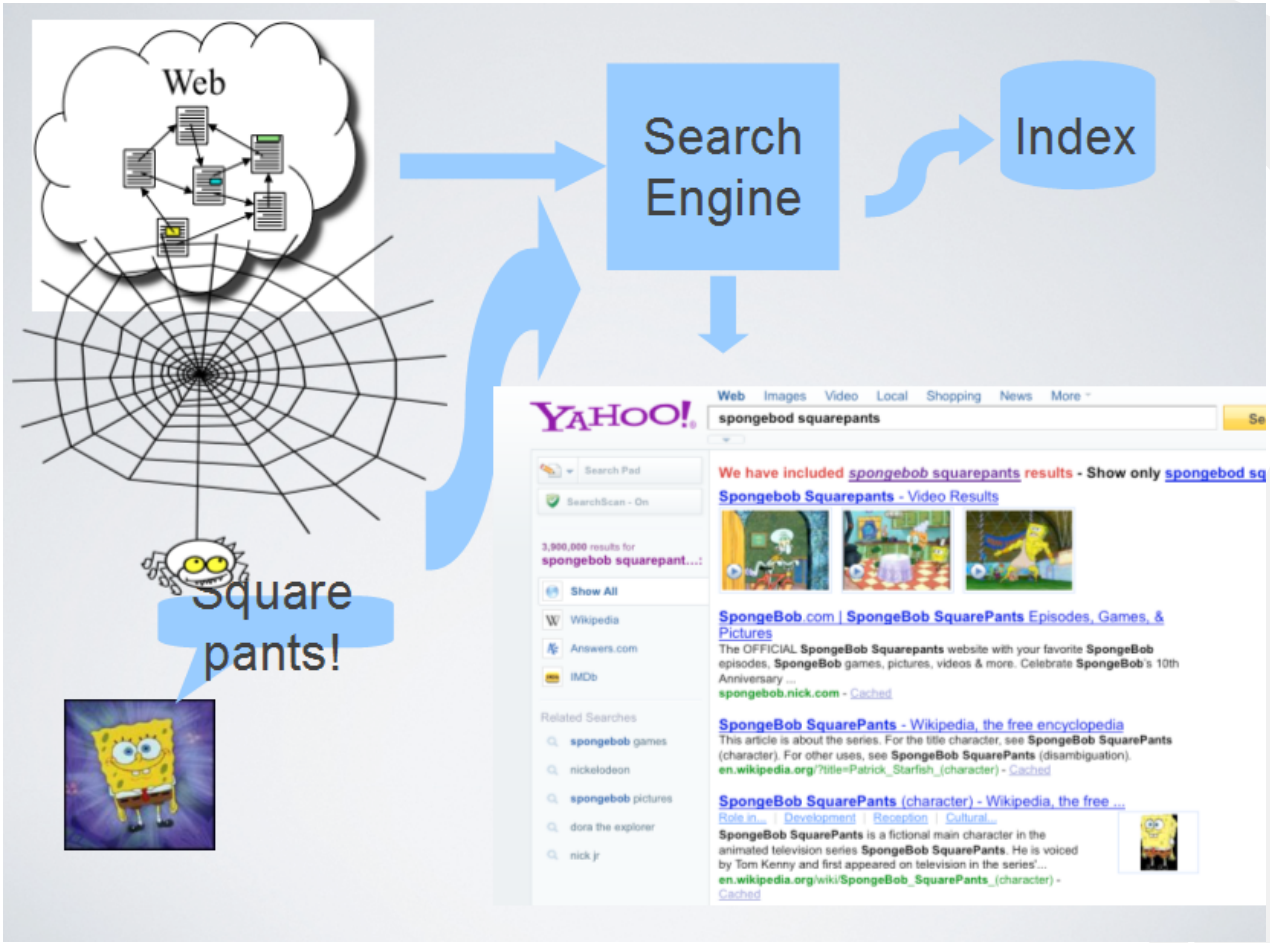
- Dinamismo
- Duplicidad
- Autoridad/Credibilidad

Ejemplo paradigmático: La Web

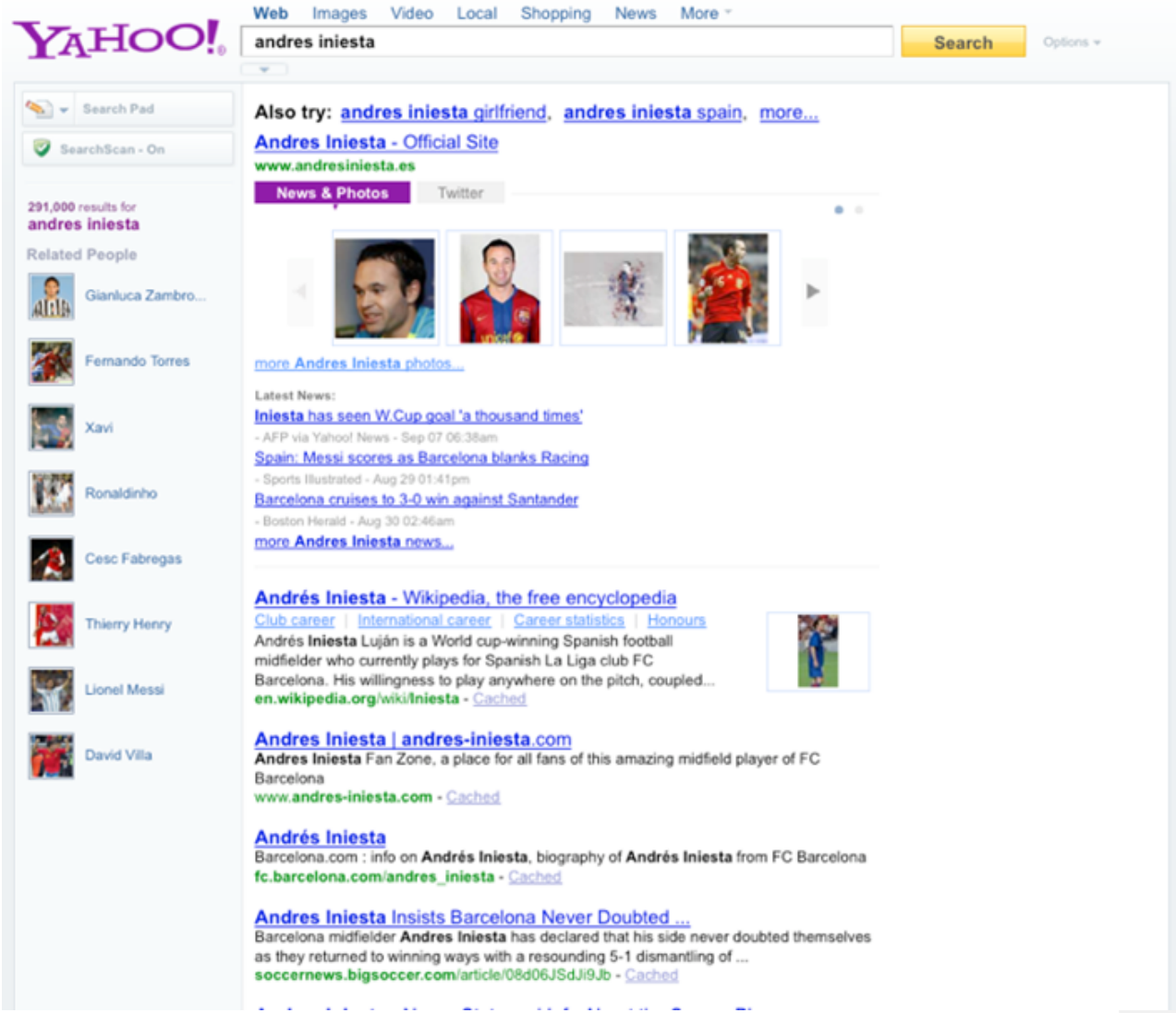
- Análisis de enlaces
- Texto de los enlaces (*Anchor text*)



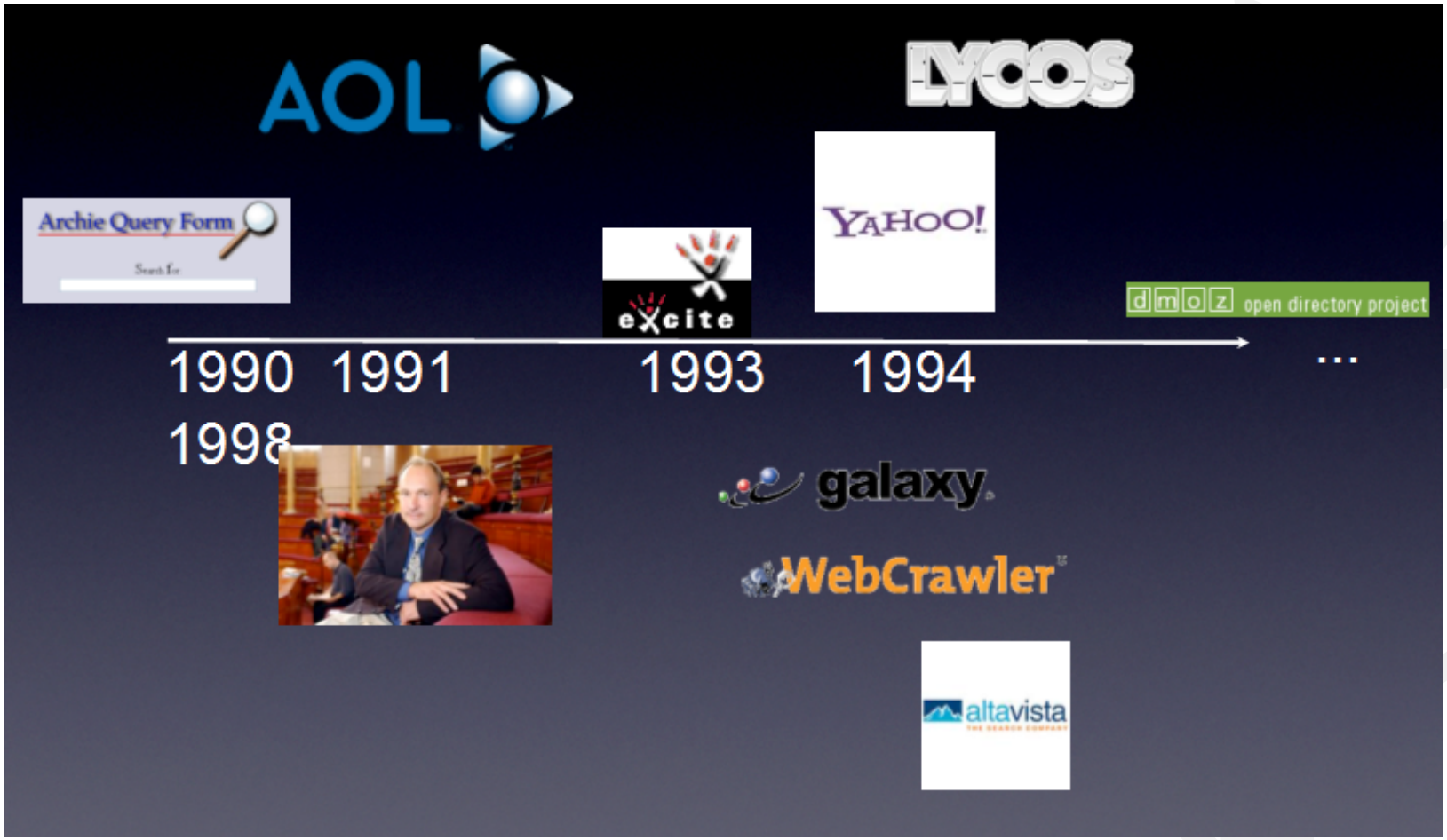
Web Search Engines



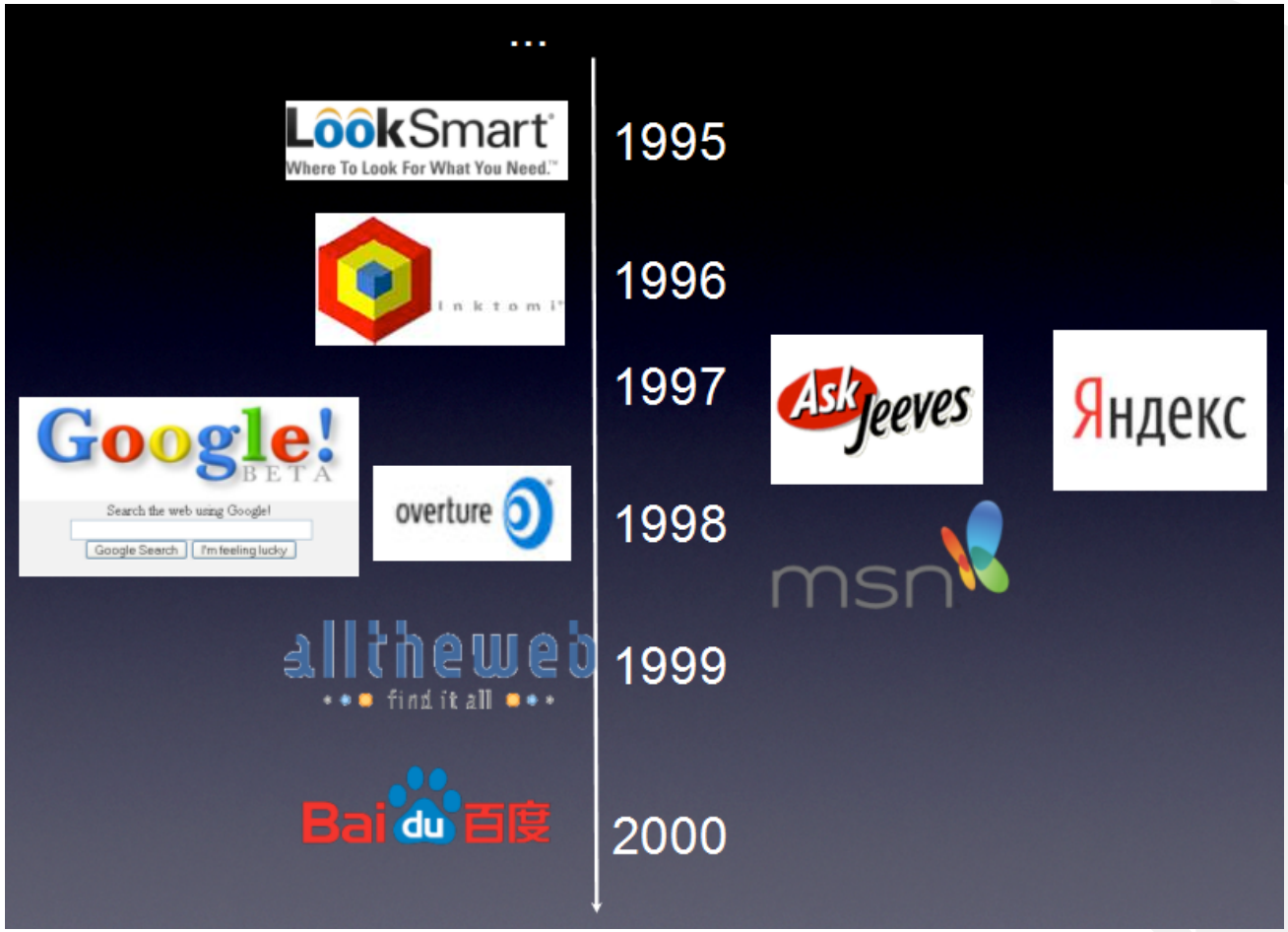
Web Search Engines (SE)



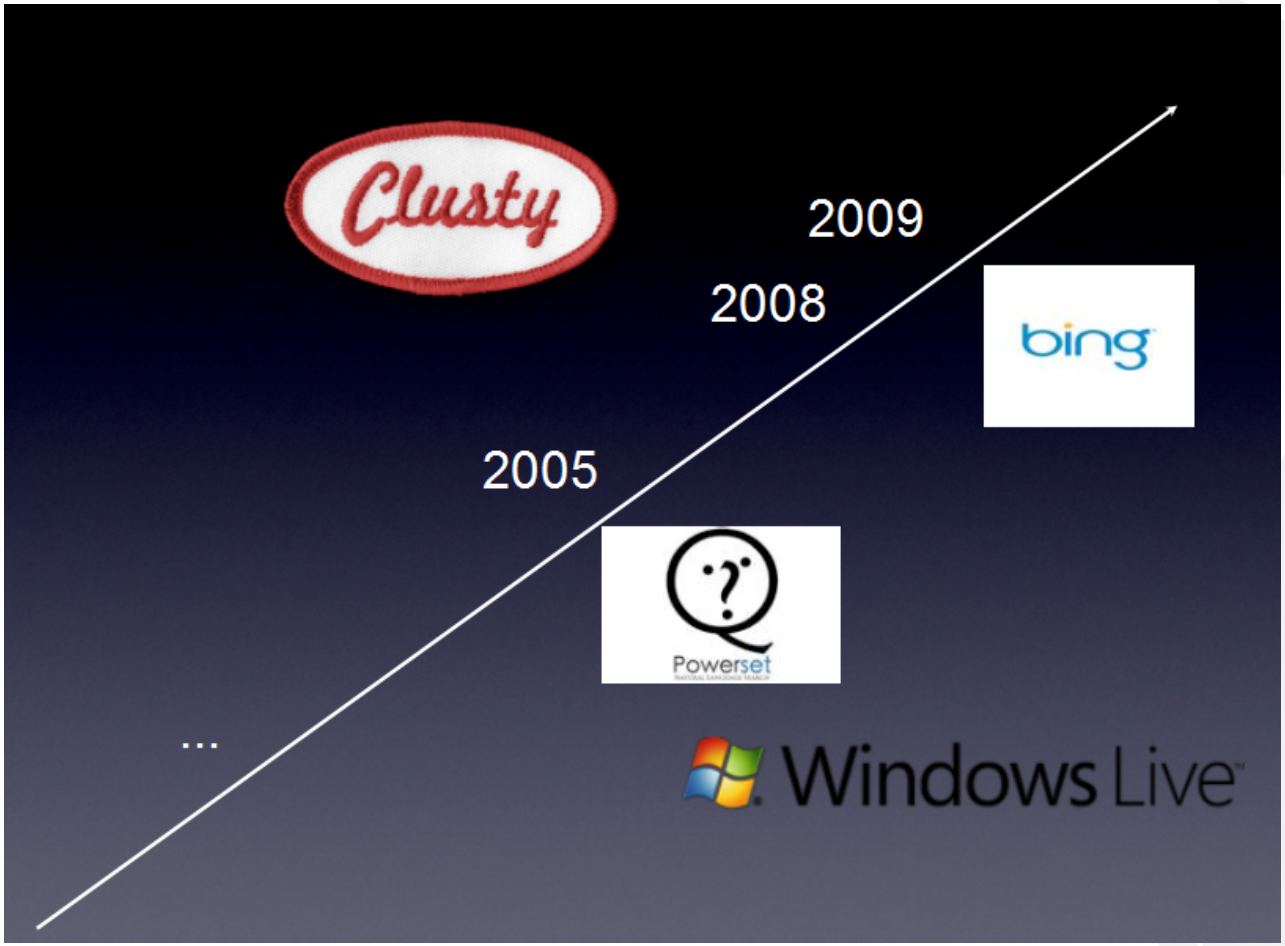
Historia de los SEs



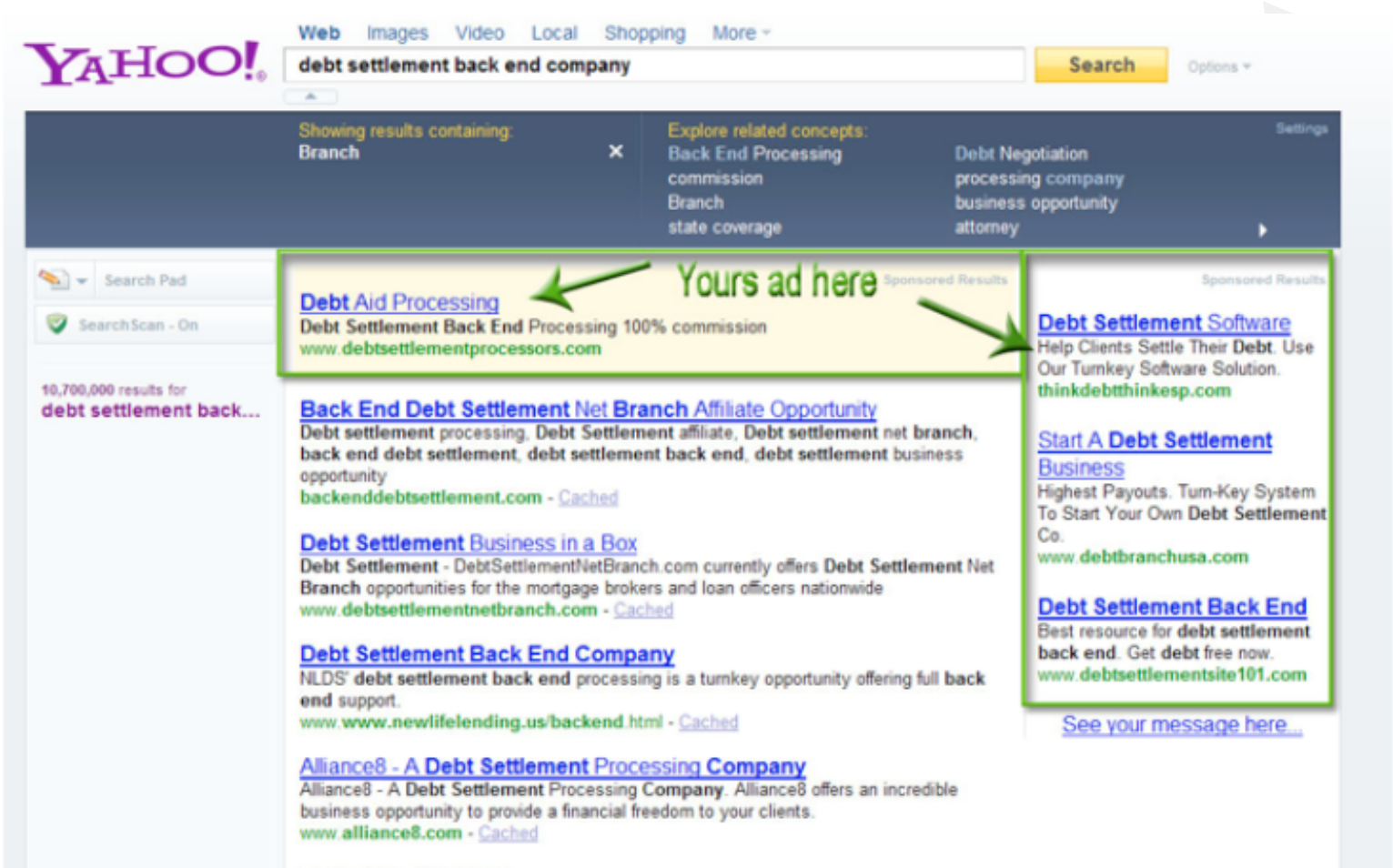
Historia de los SEs



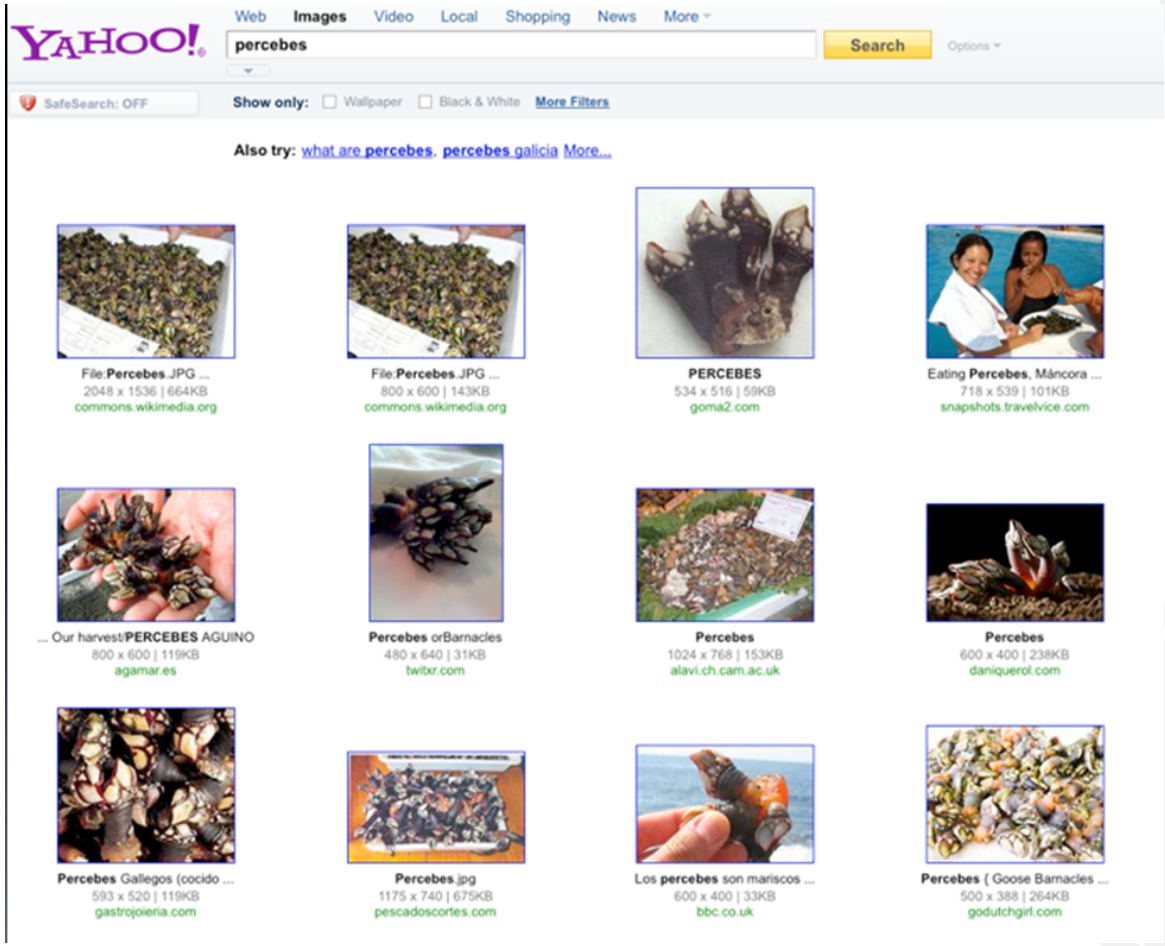
Historia de los SEs



Anuncios


YAHOO!
Web Images Video Local Shopping More
debt settlement back end company
Search
Options
Showing results containing:
Branch
Explore related concepts:
Back End Processing
commission
Branch
state coverage
Debt Negotiation
processing company
business opportunity
attorney
Settings
Search Pad
SearchScan - On
10,700,000 results for
debt settlement back...
Yours ad here
Sponsored Results
Debt Aid Processing
Debt Settlement Back End Processing 100% commission
www.debtsettlementprocessors.com
Back End Debt Settlement Net Branch Affiliate Opportunity
Debt settlement processing, Debt Settlement affiliate, Debt settlement net branch, back end debt settlement, debt settlement back end, debt settlement business opportunity
backenddebtsettlement.com - Cached
Debt Settlement Business in a Box
Debt Settlement - DebtSettlementNetBranch.com currently offers Debt Settlement Net Branch opportunities for the mortgage brokers and loan officers nationwide
www.debtsettlementnetbranch.com - Cached
Debt Settlement Back End Company
NLDS' debt settlement back end processing is a turnkey opportunity offering full back end support.
www.www.newlifelending.us/backend.html - Cached
Alliance8 - A Debt Settlement Processing Company
Alliance8 - A Debt Settlement Processing Company. Alliance8 offers an incredible business opportunity to provide a financial freedom to your clients.
www.alliance8.com - Cached
Sponsored Results
Debt Settlement Software
Help Clients Settle Their Debt. Use Our Turnkey Software Solution.
thinkdebtthinkesp.com
Start A Debt Settlement Business
Highest Payouts. Turn-Key System To Start Your Own Debt Settlement Co.
www.debtbranchusa.com
Debt Settlement Back End
Best resource for debt settlement back end. Get debt free now.
www.debtsettlementsite101.com
See your message here...

Ranking de imágenes/video

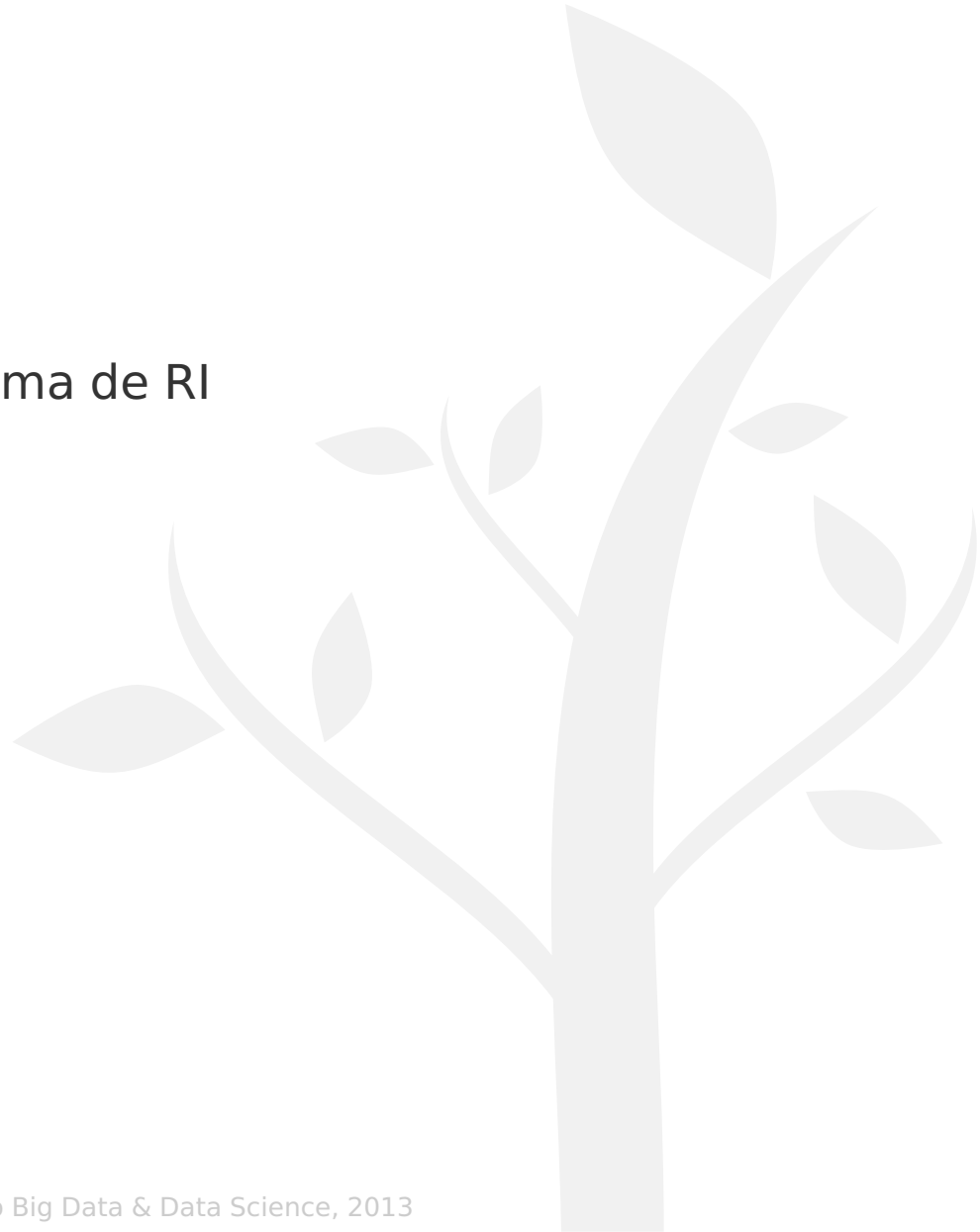


Respuesta Automática a Preguntas (QA)

Aplicaciones de la tecnología de RI

- Búsqueda de Texto/Web
- Búsqueda de anuncios
- Búsqueda de expertos
- Búsqueda multimedia
- Búsqueda de emails
- Sistemas QA
- Sistemas de Recomendación
- Búsqueda escritorio
- Búsqueda corporativa

Contenidos



Sistemas de RI

Colección de documentos

- web, artículos científicos, tests clínicos, ensayos, etc.
- Rastreo?

Rastreo

Crawler/Araña/Robot/WebBot

- Cola de URLs a visitar
- Método para recuperar y procesar recursos web
- Parser de páginas para extraer enlaces salientes
- Conexión con el indexador del SE

Rastreo

Modo de operación

1. Inicializar conj. de páginas semillas
2. Repetir
   - 2.1 Tomar URL de la cola
   - 2.2 Recuperar y parsear la pág web
   - 2.3 Extraer URLs de la pág
   - 2.4 Añadir URLs a la cola

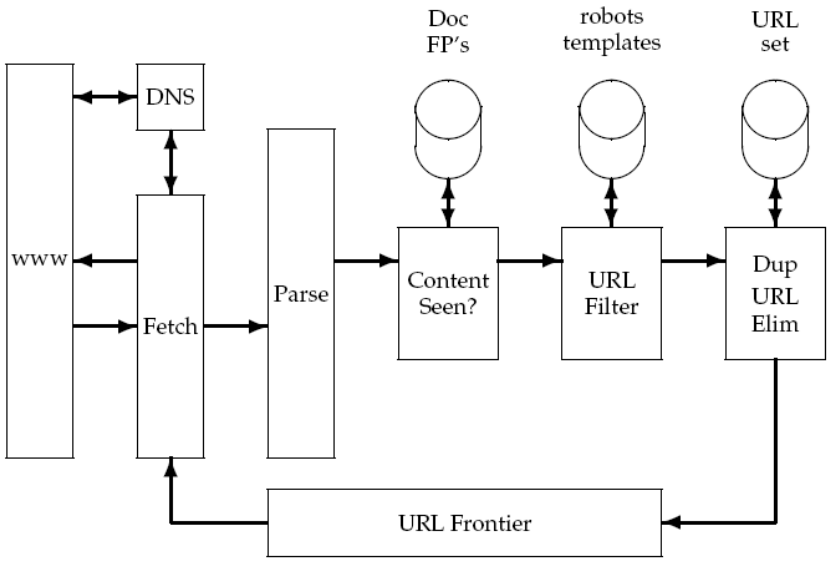


Figure 20.1 Basic crawler architecture.

Rastreo

Retos

- Escalabilidad
- Páginas spam
- Duplicidad
- *Spider traps*
- Distribuir carga (muchas arañas en paralelo)

Rastreo

Retos

- Ratio de revisita variable
- Ancho de banda
- Profundidad de rastreo
- Respetar sitios web (*politeness*) (`robots.txt`)

Indexado

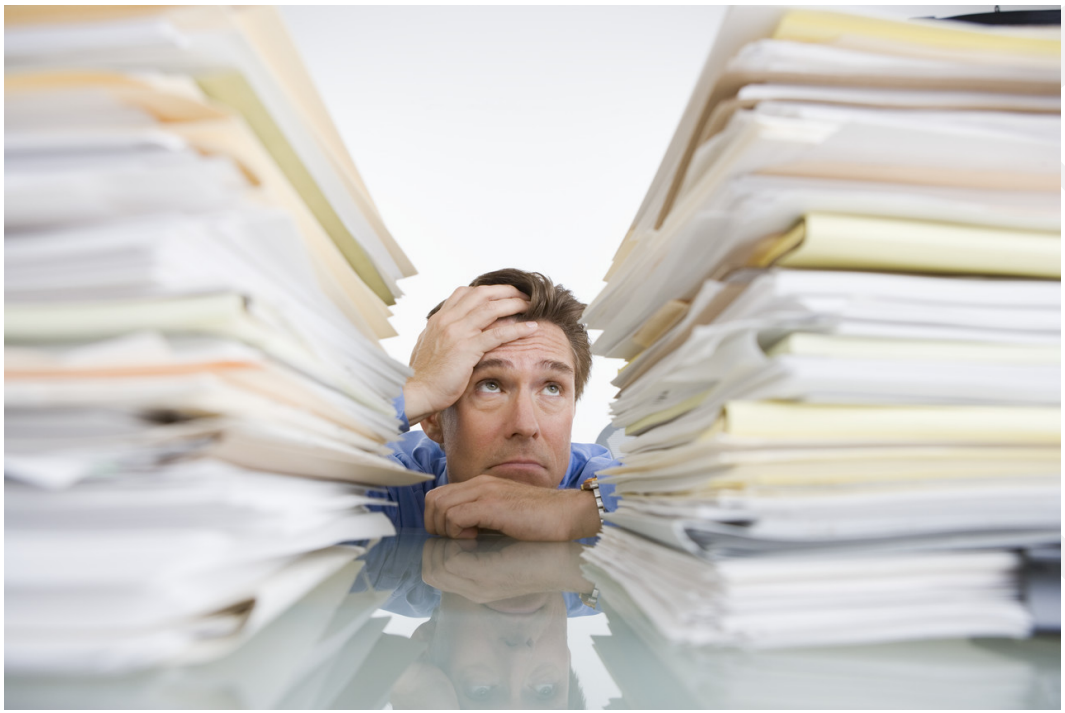
Fundamental para hacer la información accesible

- 1879: Index Medicus para catalogación de trabajos médicos
- 1960s- : MEDLINE (puede considerarse la versión electrónica de Index Medicus)

Indexado Manual

- Un humano asigna términos de indexación y atributos
- Usualmente siguiendo una terminología estandarizada (tesauro, vocabulario controlado, ...) y un protocolo específico
- Poca escalabilidad

Ej.- BD bibliográficas



Indexado Automático

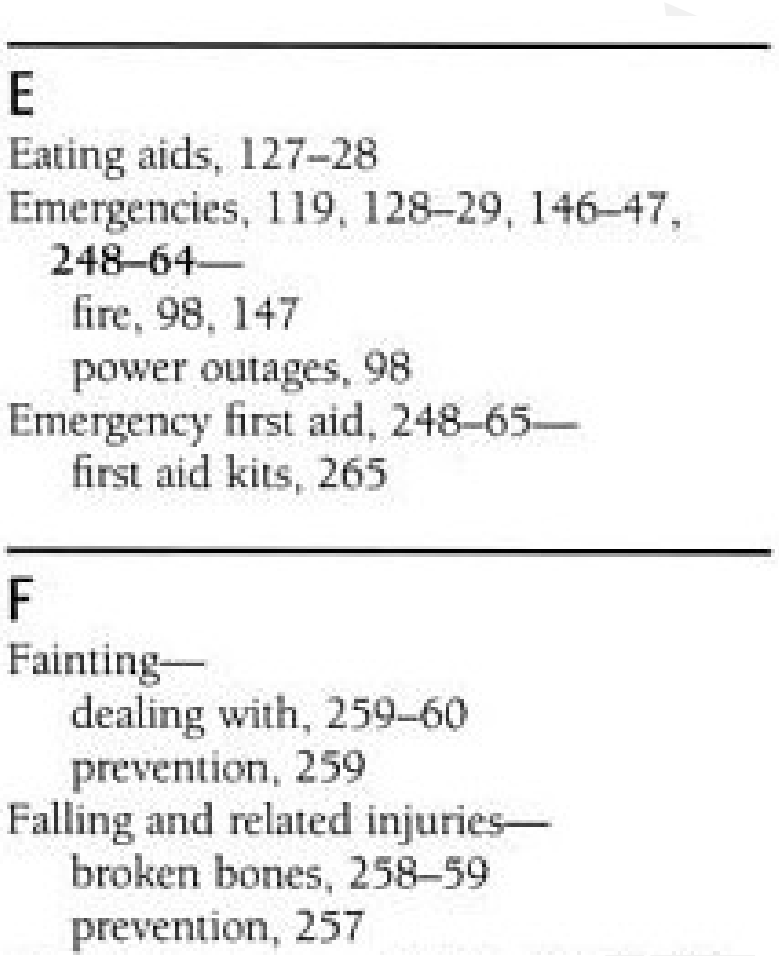
- Realizado por computadoras
- Usualmente *full-text*, limitándose a cuestiones básicas de preprocesado

Indexación Automática

- A partir de los 90 con el crecimiento de la Web
- Ficheros invertidos
- Ponderación de términos (tf/idf,...)
- Análisis de enlaces (PageRank, HITS, etc.)
- Recopilar la colección (crawling) vs colección explícita (p.e. MEDLINE)

Indexación

- Indice o fichero invertido
- vocabulario controlado o no
- metadatos
- ....



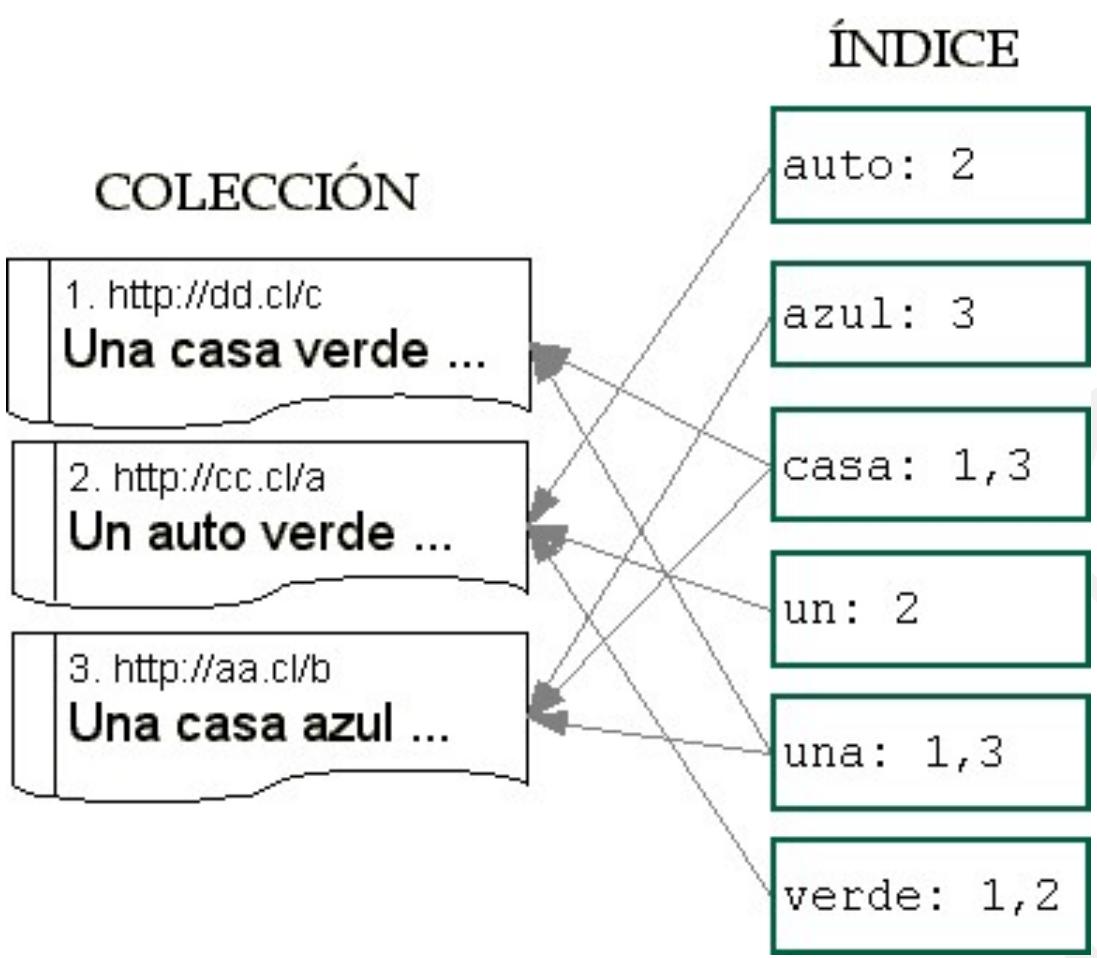
E

Eating aids, 127–28
Emergencies, 119, 128–29, 146–47, 248–64—
fire, 98, 147
power outages, 98
Emergency first aid, 248–65—
first aid kits, 265

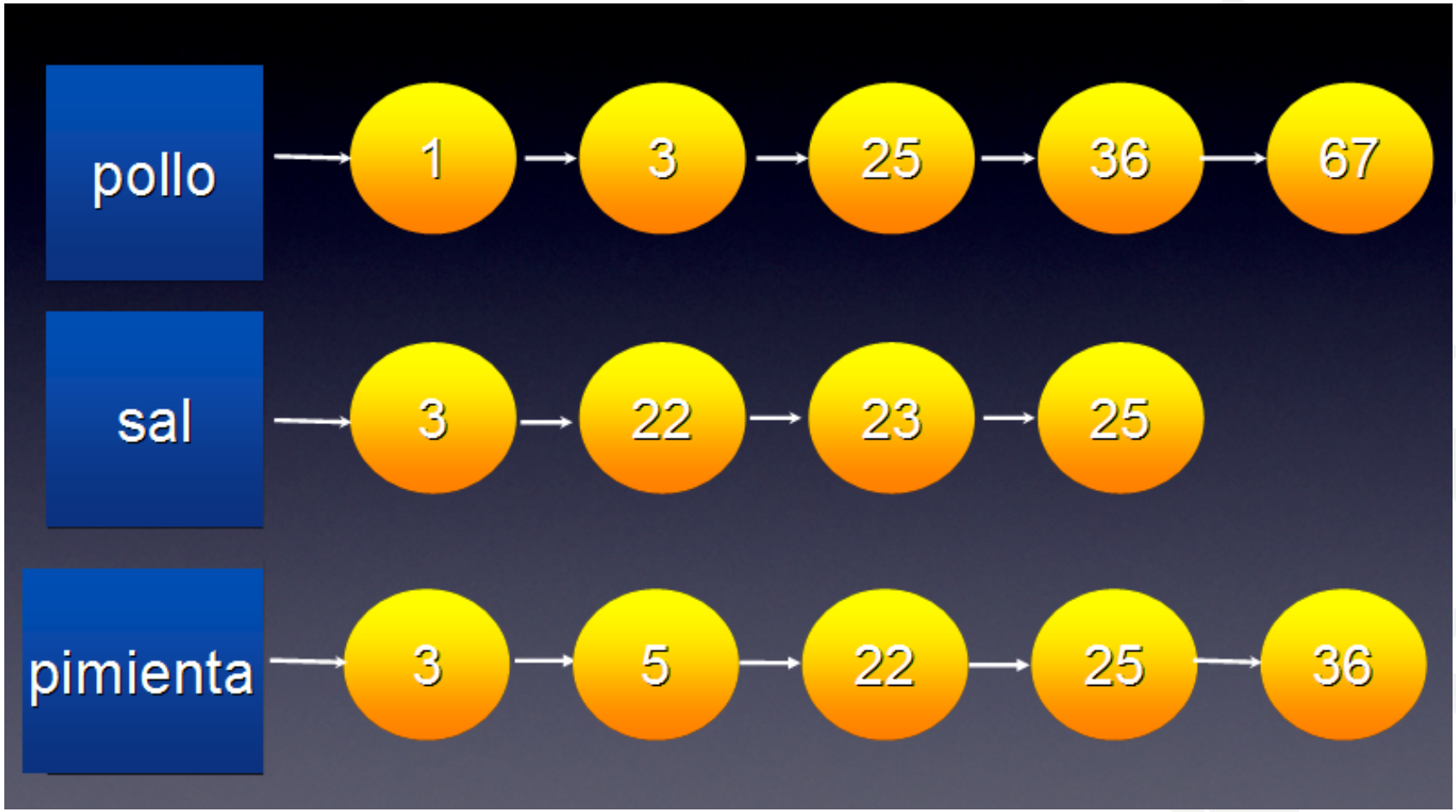
F

Fainting—
dealing with, 259–60
prevention, 259
Falling and related injuries—
broken bones, 258–59
prevention, 257

Indexación


ÍNDICE
COLECCIÓN
1. http://dd.cl/c
Una casa verde ...
2. http://cc.cl/a
Un auto verde ...
3. http://aa.cl/b
Una casa azul ...
auto: 2
azul: 3
casa: 1,3
un: 2
una: 1,3
verde: 1,2

Indice invertido



Preprocesamiento

- Stemming/Lematización

  `computer, computing, ... => comput`

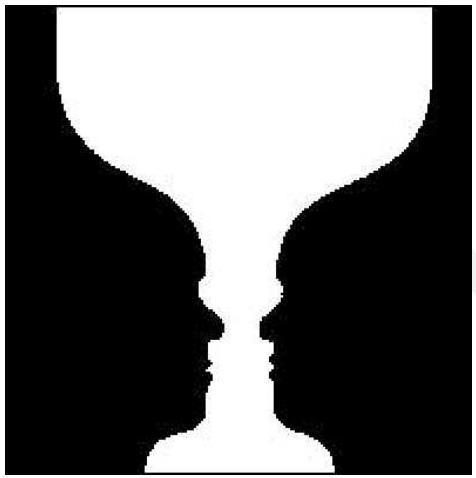
- Stopwords
- Procesamiento de Lenguaje Natural

  Part-of-Speech (POS), Reconocimiento de Entidades, etc.

Desambiguación (polisemia)

`AIDS vs hearing aids`
`EKG leads vs lead poisoning`

- Contexto, recursos terminológicos externos (p.e. Wordnet y sus synsets)



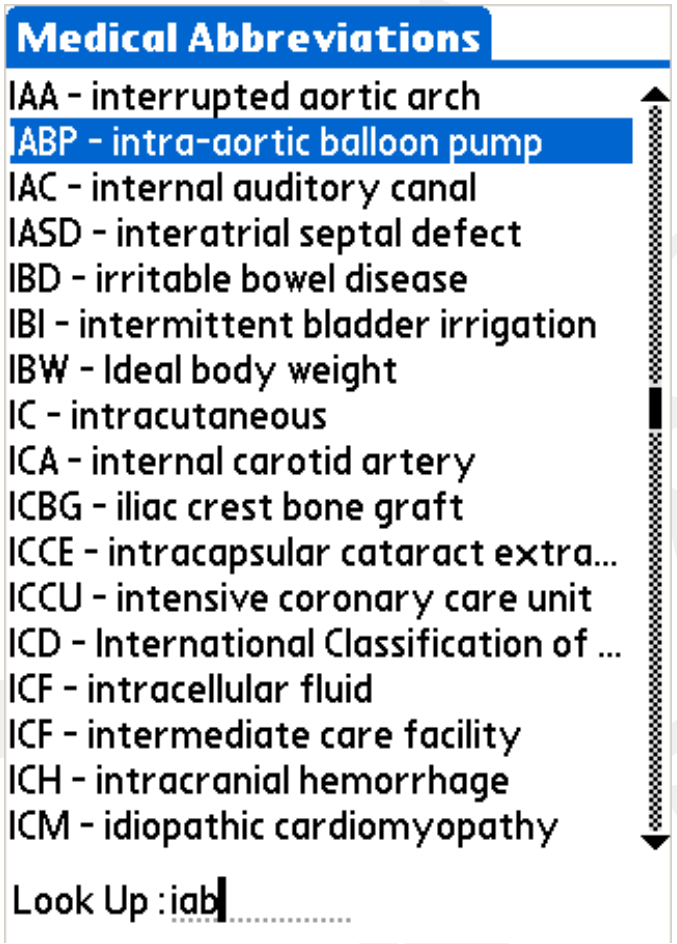
Acrónimos y nombres compuestos

Acrónimos

EPOC, RTH, VIH, ...

Nombres compuestos

Anquilomatosis cutánea, Abombamiento apical, Reticulohistiocitosis multicéntrica, ...



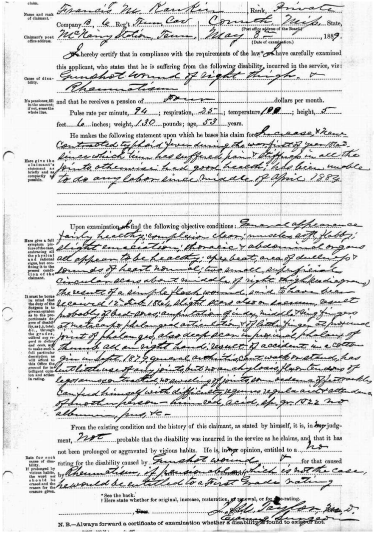
Extracción de Información

Extraer hechos o conocimiento a partir de texto libre

- Una forma de minería de textos
- Procesamiento de Lenguaje Natural (NLP)

Procesamiento de Lenguaje Natural

- Procesar la narrativa
  P.e. en informes de pacientes
  extraer síntomas, resultados de tests, condiciones físicas, etc.
- Necesaria alta precisión


I hereby certify that in compliance with the requirements of the law I have carefully examined this applicant, who states that he is suffering from the following disability, incurred in the service, viz:

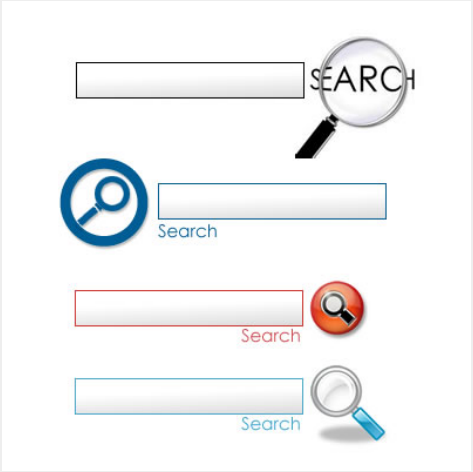
Upon examination I find the following objective conditions:

From the existing condition and the history of this claimant, as stated by himself, it is, in my judgment,

N. B.—Always forward a certificate of examination whether a disability is found to exist or not.

Sistemas de RI

Recuperación

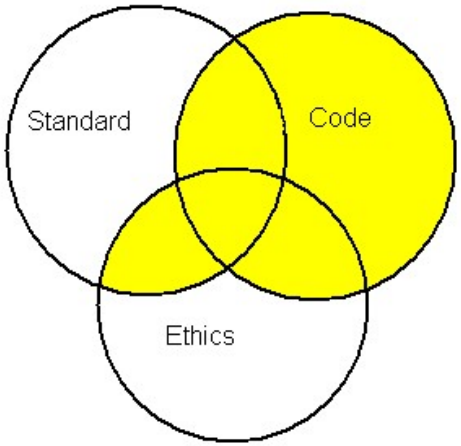


Tipo de búsqueda

Emparejamiento exacto o Búsqueda booleana (AND/OR/NOT)

```
propanolol AND hypertension
```

- Dificultad de comprensión de las conectivas booleanas
- Difícil controlar el tamaño del conjunto de salida
- Estimación binaria de la relevancia

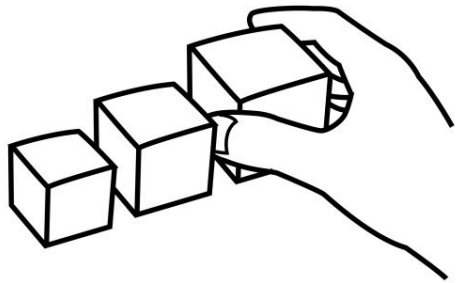


Emparejamiento parcial

Ranking

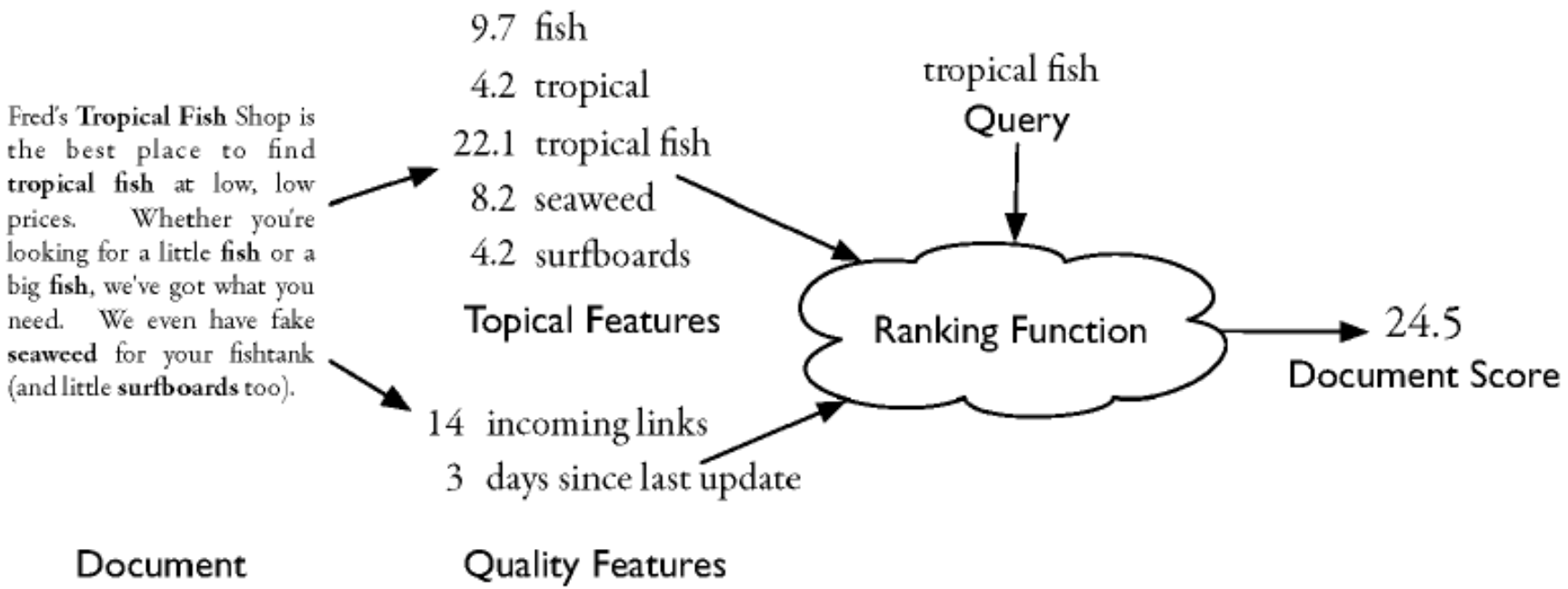
- Consulta en lenguaje natural libre
- Ponderación de términos (tf/idf y sus muchas evoluciones)
- Funciones de ranking

  Modelos probabilísticos (BM25, Language Models, etc)



Ranking

Factores dependientes de la consulta + factores independientes



Texto de los enlaces (*Anchor Text*)

IBM

IBM manufactures and sells computer services, hardware, and software. Also provides financing services in support of its computer business.

www.ibm.com

Análisis de Enlaces

PageRank (Google)

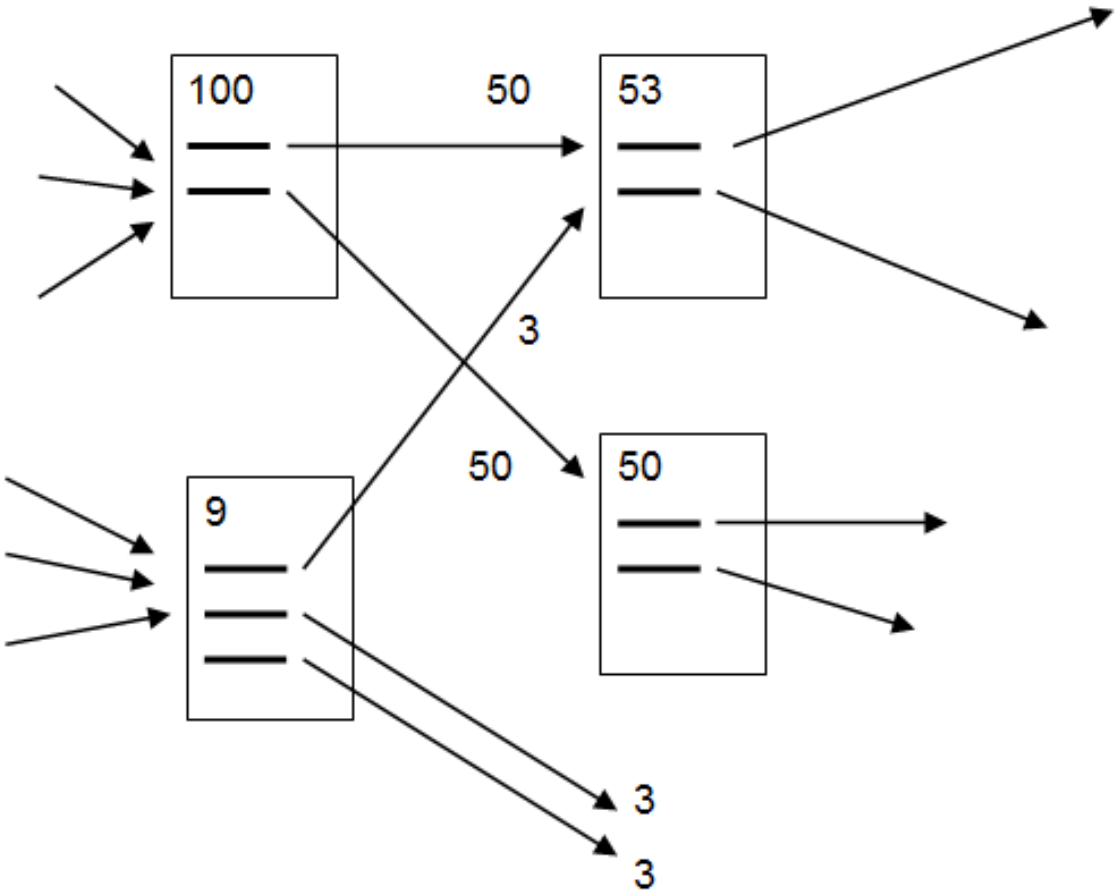
- Sistema *democrático* de votación
  1 enlace entrante $\rightarrow$ 1 voto
- Simple cuenta de enlaces no es suficiente
- El *PageRank* de una página
  Medida objetiva de importancia

Análisis de Enlaces

PageRank (Google)

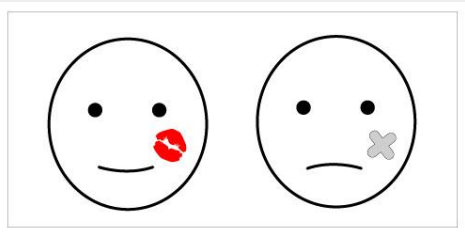
- Si $T_1, T_2, \cdots, T_n$ apuntan a una pág. A
  $PR(A) = (1-d) + d \cdot (\frac{PR(T_1)}{C(T_1)} + \cdots + \frac{PR(T_n)}{C(T_n)})$
  $C(P)$, número de enlaces que tiene la página P
  $d \in [0, 1]$, factor de escalado
- PR crece con el núm. de *inlinks*
- PR crece con la calidad de las págs que apuntan a A
  (ej.- pág. que figura en el directorio de Yahoo)

PageRank (Google)



Mejorar la búsqueda

Realimentación de relevancia



Selección de términos y Expansión de consultas

Automática o asistida

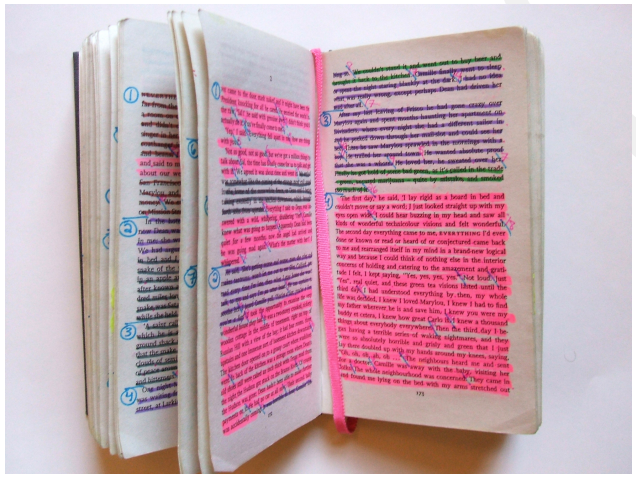
Más servicios de búsqueda

Operadores de proximidad

```
colon ADJ5 cancer
```

Recuperación de pasajes

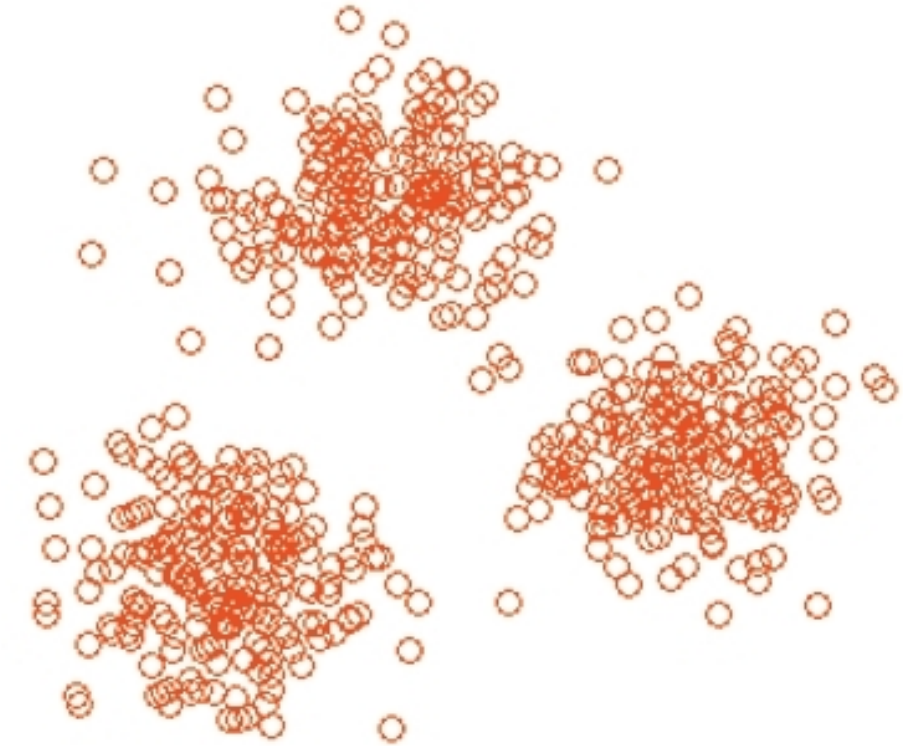
- discurso (párrafos, oraciones, etc.)
- semánticos (automatizar el troceado conceptual de los textos)
- de ventana



Agrupamiento

Clustering

- grupos no predefinidos
- descubrir las clases de los ejemplos



Filtrado

- Flujo de docs (p.e. stream de noticias)
- Esencialmente técnicas estándar de RI
- Perfil de interés $\leftrightarrow$ Consulta
- Adaptativo o por lotes (batch)

Generación Automática de Resúmenes

- Extracts vs Abstracts
- Grado de reducción, Informativos y bien formados
- Orientados a consulta (p.e. web snippets) o no

Varicela: MedlinePlus enciclopedia médica
www.nlm.nih.gov/medlineplus/spanish/ency/article/001592.htm
8 Feb 2011 – La **varicela** se puede contagiar muy fácilmente a otras personas. Usted puede contraerla tocando los líquidos de una ampolla de **varicela** o si ...
↳ Vista de cerca - Lesión en la pierna - Lesiones en el tórax por varicela - Varicela

Recuperación con Lenguas Cruzadas

CrossLingual Information Retrieval (CLIR)

- p.e para recuperar datos de ensayos clínicos no escritos en nuestra lengua de origen
- p.e. para propósitos educacionales: aprender la terminología médica en Inglés

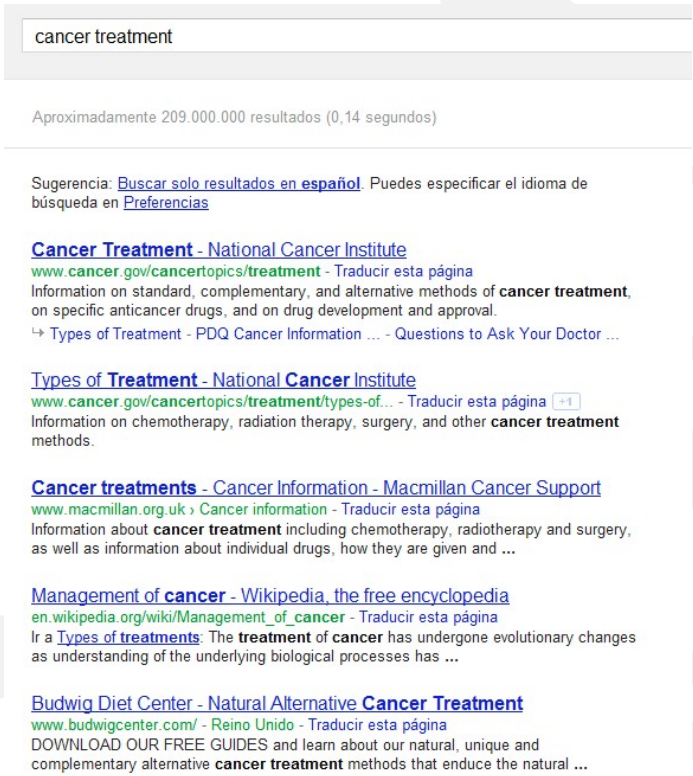
Contenidos

Frontiers, Challenges, and Opportunities for Information Retrieval (SWIRL 2012)

Temas comunes

Not just a ranked list

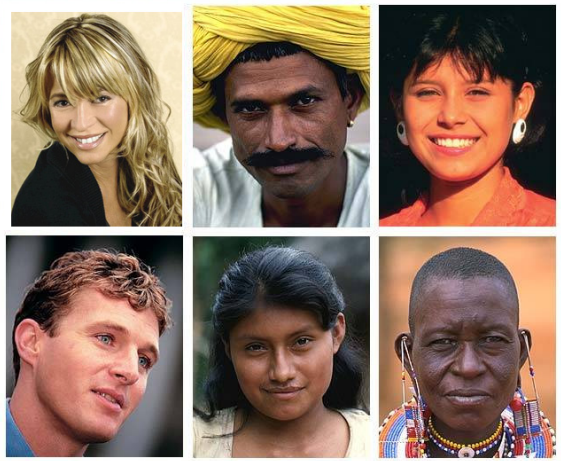
Move beyond the classic *single adhoc query and ranked list* approach, considering richer modes of querying, models of interaction, and approaches to answering



Temas comunes

Help for users

Ways that IR technology can be extended to support users more broadly, including ways to bring IR to inexperienced, illiterate, and disabled users



Temas comunes

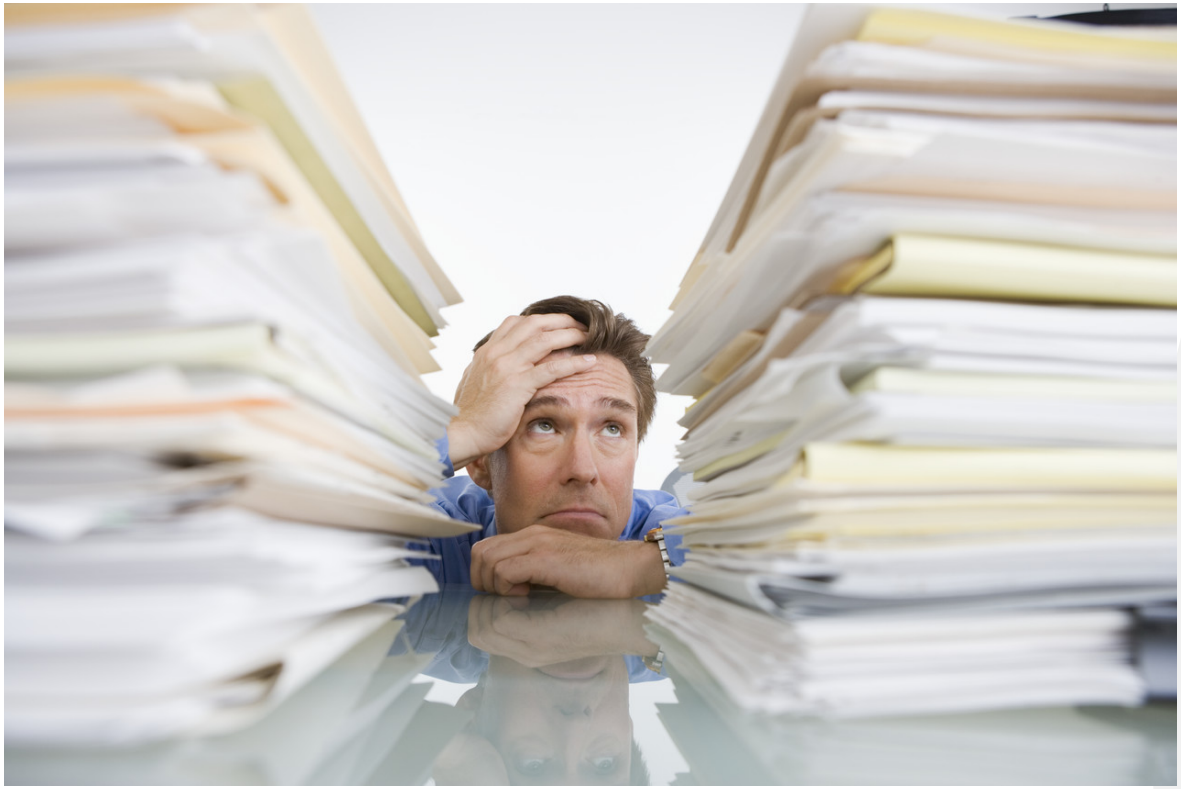
Capturing context

Ways to incorporate what is happening with and around a user to affect querying and result presentation. In particular, this theme treats people using search systems, their context, and their information needs as critical aspects needing exploration

Temas comunes

Information, not documents

Push Information Retrieval research beyond document retrieval and into more complex types of data and more complicated results



Temas comunes

Domains

Consider information that is not simply text and that has not been thoroughly explored by IR research so far – data with restricted access, collections of *apps,* and richly connected workplace data

Temas comunes

Evaluation

A perennial issue in Information Retrieval, evaluation remains important, particularly as the field expands into new challenges. This theme includes topics that require or suggest new techniques for evaluation as well as those that need evaluation in the context of new challenges

		Not just a ranked list	Help for users	Capturing context	Not just documents	New domains	Evaluation
4.1	Conversational Answer Retrieval	X			X		
4.2	Empowering Users To Search and Learn	X	X	X			
4.3	Finding What You Need with Zero Query Terms	X		X			
4.4	Mobile Information Retrieval Analytics			X	X	X	X
4.5	The Structure Dimension				X	X	
4.6	Understanding People in Order to Improve I(R) Systems			X			X
5.1	Abstracting Information Retrieval Evaluation						X
5.2	Adapting to Various Sites, Tasks and Contexts			X			
5.3	Axiometrics – Foundations of Evaluation Metrics in IR						X
5.4	Before and After the Mobile Query			X			
5.5	Community Evaluation Service						X
5.6	Exploring the Intersection of Social and Algorithmic Search						
5.7	Getting Your Life Back: … Personal Data					X	
5.8	*Information* Retrieval				X		
5.9	IR4ALL: Addressing … Divides to Search		X				
5.10	Information Retrieval for the Ecosystem of Apps			X		X	
5.11	Information Seeking Stage Aware Search	X	X	X			
5.12	Protecting Users' Privacy in Search			X			
5.13	Search Among Secrets					X	
5.14	Simulation of Interaction	X					X
5.15	Spoken Information Retrieval		X		X		
5.16	Super Models of Information Retrieval Interaction	X					
5.17	Supporting Complex Search Tasks	X					
5.18	Time Changes Everything	X			X	X	
5.19	Understanding and Evaluating Rich Aggregated Answers	X			X		X
5.20	Understanding Opinion Engineering					X	
5.21	Understanding Search in the Workplace			X		X	

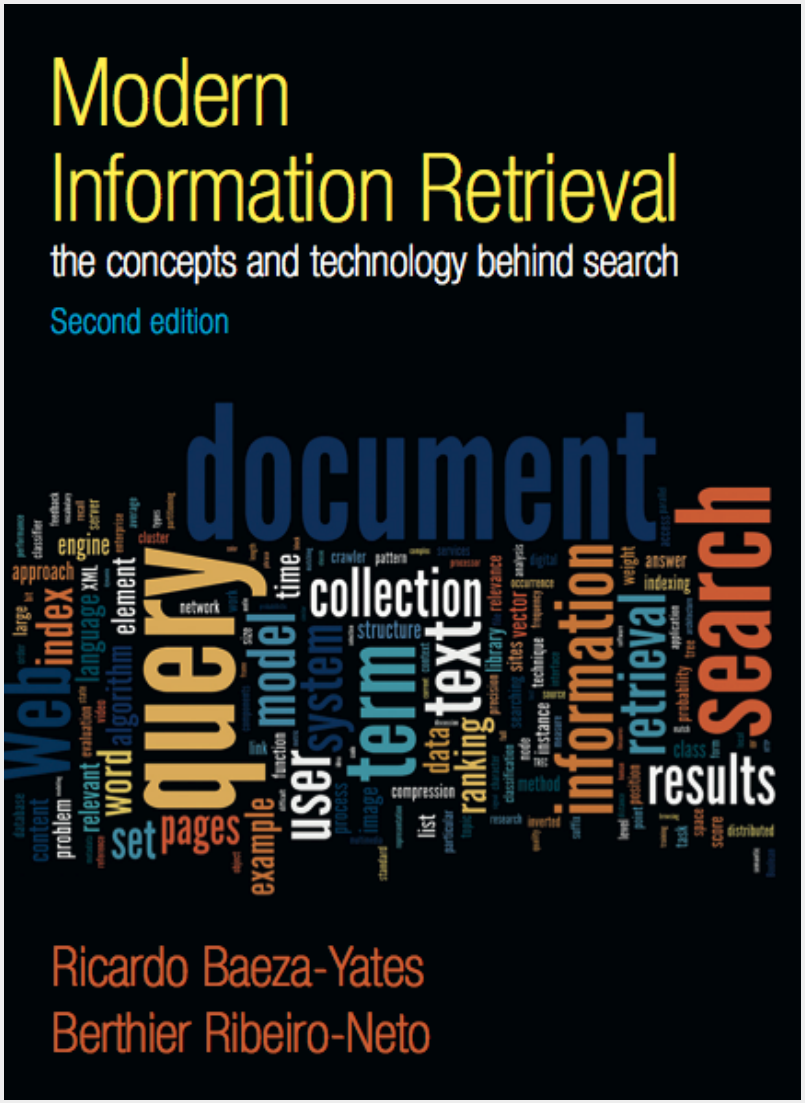
http://www.cs.rmit.edu.au/swirl12/

SWIRL'12: The Second Strategic Workshop on Information Retrieval. Lorne (Australia)

- James Allan, Bruce Croft, Alistair Moffat, and Mark Sanderson (eds.), "Frontiers, Challenges, and Opportunities for Information Retrieval: Report from SWIRL 2012."SIGIR Forum, 46(1):2-32, June 2012.
- An annotated list of recent papers: key challenges and developments in the IR area.

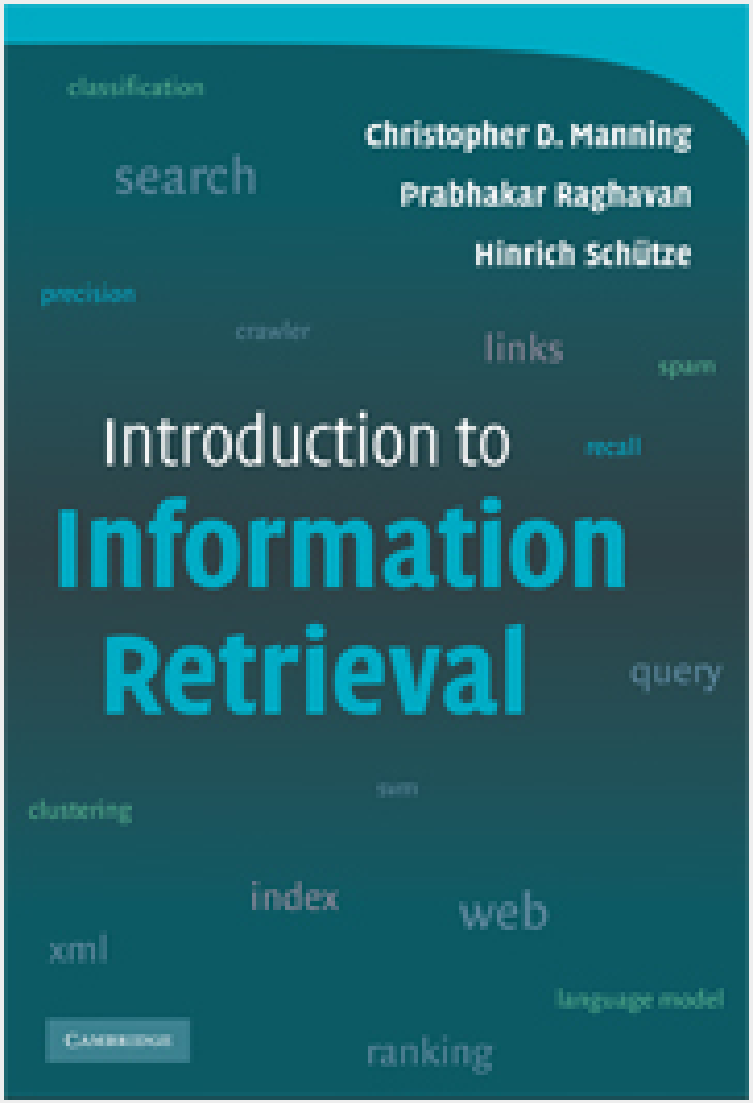
Referencias bibliográficas

www.mir2ed.org

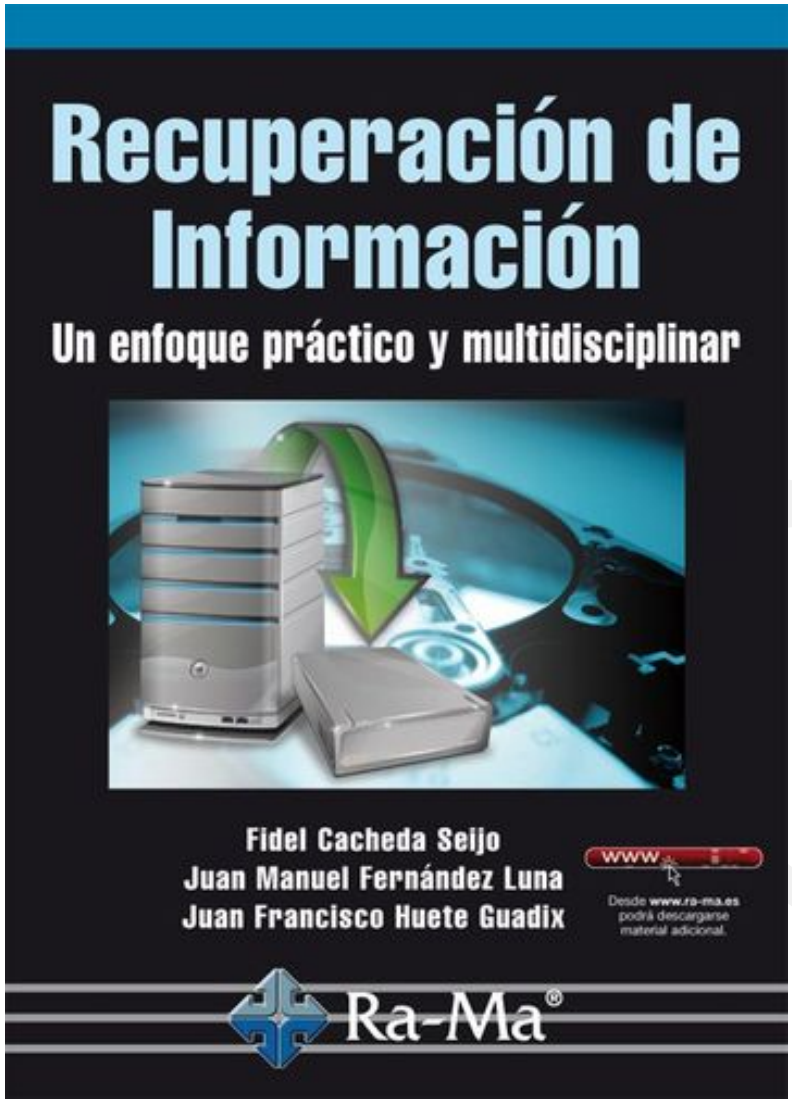


Referencias bibliográficas

nlp.stanford.edu/IR-book/



Referencias bibliográficas



Más info y contacto

http:\\www.gsi.dec.usc.es\ir
david.losada@usc.es