Encontrar documentos a través de las palabras

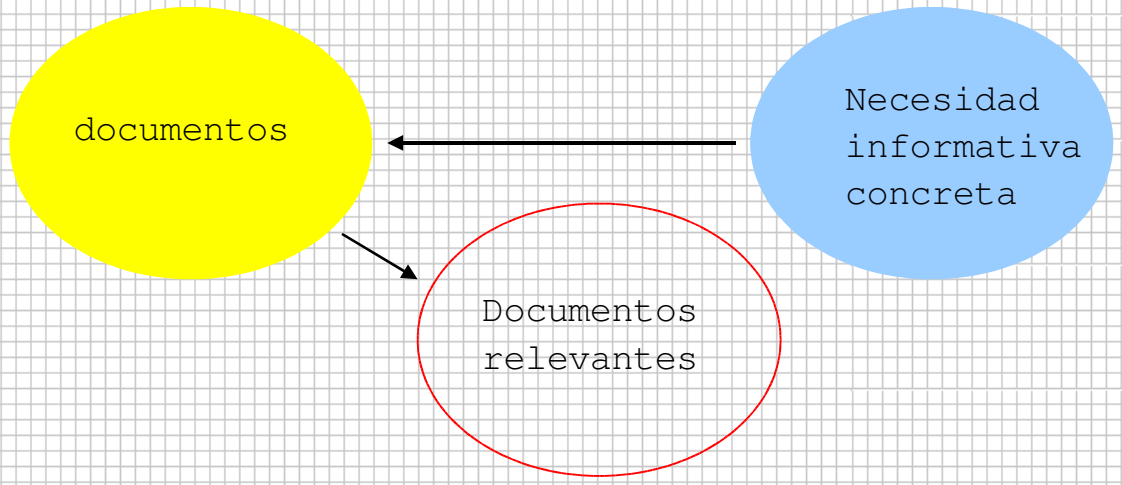
Carlos G. Figuerola
Universidad de Salamanca
Grupo REINA
http://reina.usal.es

El problema de la RI

Crecimiento exponencial de la documentación

Necesidad de seleccionar los documentos que satisfagan las necesidades informativas concretas

El problema se centra en la búsqueda por temas o contenidos



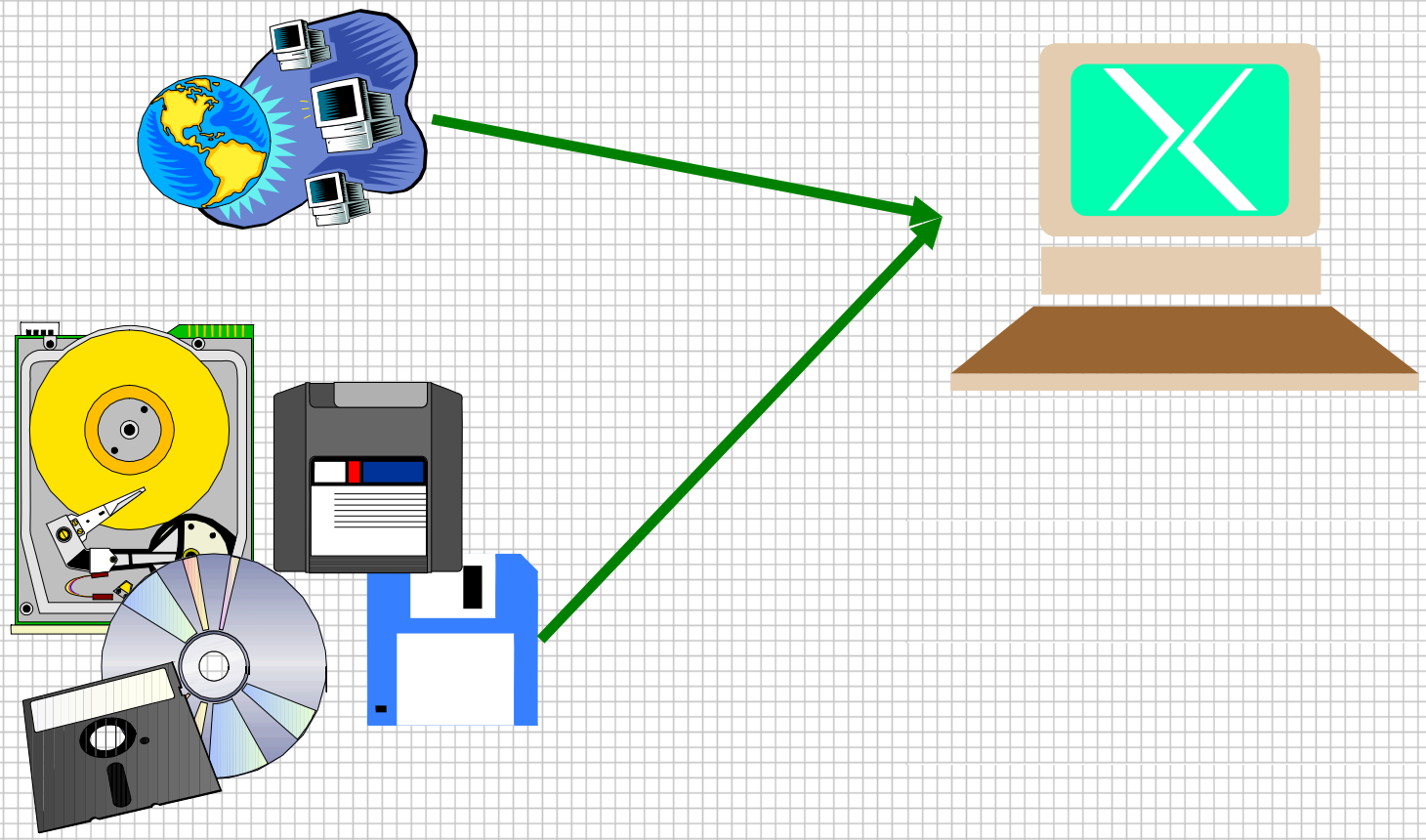
La indización manual

Inconvenientes

es muy costosa en tiempo y trabajo

inconsistencias inevitables entre indizadores

Documentos electrónicos



La indización automática

Inconvenientes

Información pobremente estructurada

Solución simple

Búsqueda de subcadenas, mediante ficheros invertidos u otros sistemas.
Utilización de operadores booleanos y de proximidad

Búsquedas de subcadenas

Problemas

sinonimia y polisemia

dificultad para el usuario

los documentos recuperados son todos igual de relevantes

Modelo vectorial

Cada documento es representado como un vector o lista de términos

Cada término tiene un **peso** que indica su importancia dentro de cada documento

Las necesidades de información del usuario se formulan en lenguaje natural

se representan también como una lista o vector de términos, y cada término tiene también un peso que indica su importancia

Modelo vectorial (ejemplo)

<DOC>
<CLAVE>DTT001-0267</CLAVE>
<TITULO> Configuración de redes locales en CD-ROM.
</TITULO>
<RESUMEN> Analiza la configuración del equipo físico y lógico para la integración de una red de área local de más de 50 ordenadores con aplicaciones en CD-ROM. Incluye una reseña sobre la oferta de productos en el mercado. Finalmente se concluye con el proyecto llevado a cabo en la Universidad Carlos III de Madrid</RESUMEN>
</DOC>

Modelo vectorial (ejemplo)

Término	doc
configuracion	267
redes	267
locales	267
CD-ROM	267
analiza	267
configuracion	267
equipo	267
....	....
archivo	268
equipo	268

ORDENAR
Y
CALCULAR
PESOS
→

Término	Doc	Frec	Peso
analiza	267	1	0.1
archivo	268	1	0.3
CD-ROM	267	2	0.6
configuracion	267	2	0.7
equipo	267	1	0.2
equipo	268	1	0.2
locales	267	1	0.6
redes	267	1	0.4
.....	.....	....	

¿Cómo se calculan los pesos?

- diversos sistemas de estimación
- presunciones básicas:
  - un término tiene menor poder discriminatorio cuanto más frecuente es en la colección de documentos
  - un término es representativo de un documento si aparece muchas veces en ese documento

El modelo vectorial. Pesos

- tres componentes en el cálculo de los pesos
  - la frecuencia del término en eldocumento
  - el IDF (Inverse Document Frequency)
  - un factor de normalización

$$Peso(T_i D_k) = \frac{frec_{T_i D_k} \times IDF_{T_i}}{normalizador_{D_k}}$$

Consultas

- las necesidades de información se formulan en lenguaje natural (consultas)

- se tratan igual que los documentos:
  - se representan mediante una lista de términos (vectores)
  - los términos tienen también pesos que expresan la importancia de cada término en la consulta

Resolución de Consultas

- Se estima la similitud entre el vector de la consulta y cada uno de los vectores de los documentos
- Existen diversas funciones matemáticas que permiten calcular la semejanza entre dos vectores
- El resultado de comparar dos vectores es un coeficiente que expresa el grado de parecido entre ambos

Resolución de Consultas. Ejemplo

Pesticidas en alimentos para bebes
Encontrar noticias sobre pesticidas en alimentos para bebes.
Los documentos relevantes proporcionan información sobre el descubrimiento de pesticidas en alimentos para bebes. Se informa sobre diferentes marcas, supermercados y compañías que ofrecieron alimentos para bebes que contenían pesticidas. Se discuten también medidas contra la contaminación de alimentos para bebes con pesticidas.

Resolución de Consultas. Ejemplo

Doc.Nº	Simil	Título
42516	28.00	BANCO HAMBRE PRODUCTOS DE DESECHOS ALIMENTAN A 50.000 PERS
172743	26.00	UE-AGRICULTURA GREENPEACE DENUNCIA SUBVENCIONES EXPORTACION
55464	20.00	RFA-ALIMENTOS MAS POTITOS PROCEDENTES DE ESPAÑA CON RASTROS
134812	19.00	MEDIO AMBIENTE AGRICULTURA ECOLOGICA OCUPA 0,1 % TIERRAS CULT
83220	18.00	EEUU-ALIMENTACION NIÑOS Y JOVENES DE EEUU EXPUESTOS AL CANCER
56832	18.00	RFA-ALIMENTOS PARLAMENTO SE OCUPARA DE POTITOS CONTAMINADOS
49748	18.00	HAITI-EMBARGO DERECHISTAS PROTESTAN POR LLEGADA BUQUE FRA
121278	17.00	LA CARNE PICADA Y EL POLLO SON LOS ALIMENTOS MAS CONTAMI
133491	16.00	PRECIOS-REACCIONES ECONOMIA ESPERA DESCENSO INFLACION PRO
46940	16.00	OBESIDAD-TEORIAS DESMIENTEN OBESIDAD SEA CAUSA INGESTION EX
13245	16.00	CHINA-PESTICIDAS MAS DE 10.000 MUERTOS A CAUSA PESTICIDAS VEN
56184	15.00	RFA-ALIMENTACION FISCALIA SE OCUPA DE POTITOS CONTAMINADOS
178697	14.00	SUIZA-ALIMENTACION AUMENTA EN UN 50 POR CIENTO CONSUMO ALIM
175502	14.00	PRESENTADA MADRID FUNDACION "BANCO DE ALIMENTOS DE ESPAÑA"
126904	14.00	JUBILADOS Y AMAS DE CASA AL FRENTE DE UN BANCO DE ALIMENTOS
119421	14.00	MEXICO-NIÑOS ALIMENTO DIARIO PARA 64.000 NIÑOS INDIGENAS MEXIC
108094	14.00	ARGENTINA-MEDIO AMBIENTE DENUNCIAN ENTIERRO CLANDESTINO DE
85423	14.00	BRASIL-INFLACION GOBIERNO SE PREPARA PARA COMBATIR ESCASEZ

Evaluación

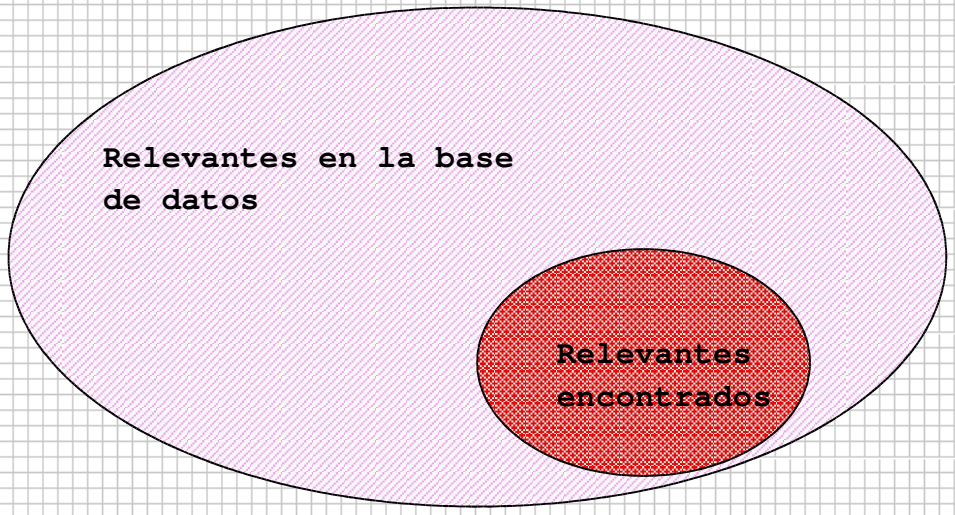
- Después de planteada la necesidad informativa y de obtener los documentos, es necesario evaluar si éstos se corresponden con la necesidad informativa

- Aspectos:
  - Velocidad de respuesta, presentación de la salida, interfaz de usuario
  - Efectividad de la recuperación: precisión, exhaustividad

    La mayor parte de medidas están determinadas por los resultados comparativos entre documentos recuperados y documentos relevantes para una consulta dada.

Evaluación

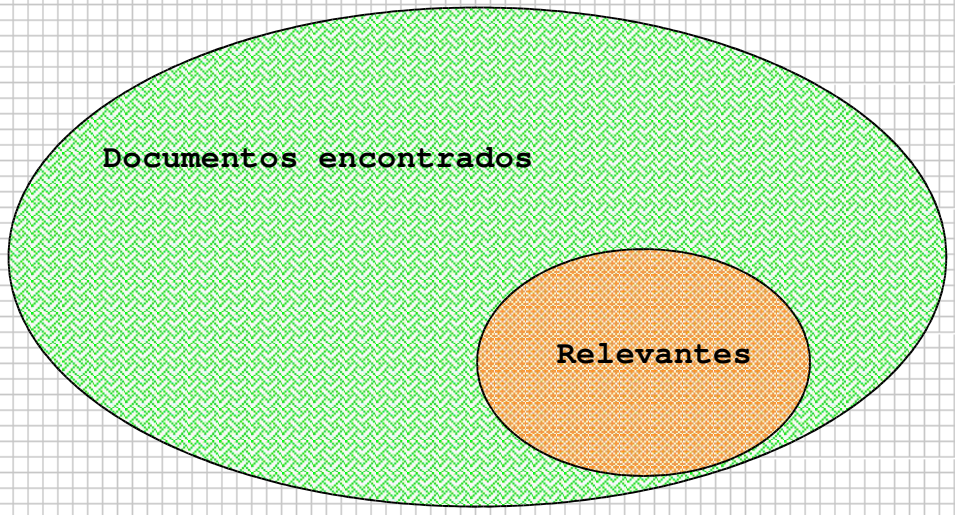
Exhaustividad:
Proporción de documentos relevantes encontrados del total de documentos relevantes en la base de datos



(sólo calculable en bases de datos experimentales)

Evaluación

Precisión:
Proporción de documentos relevantes entre los recuperados

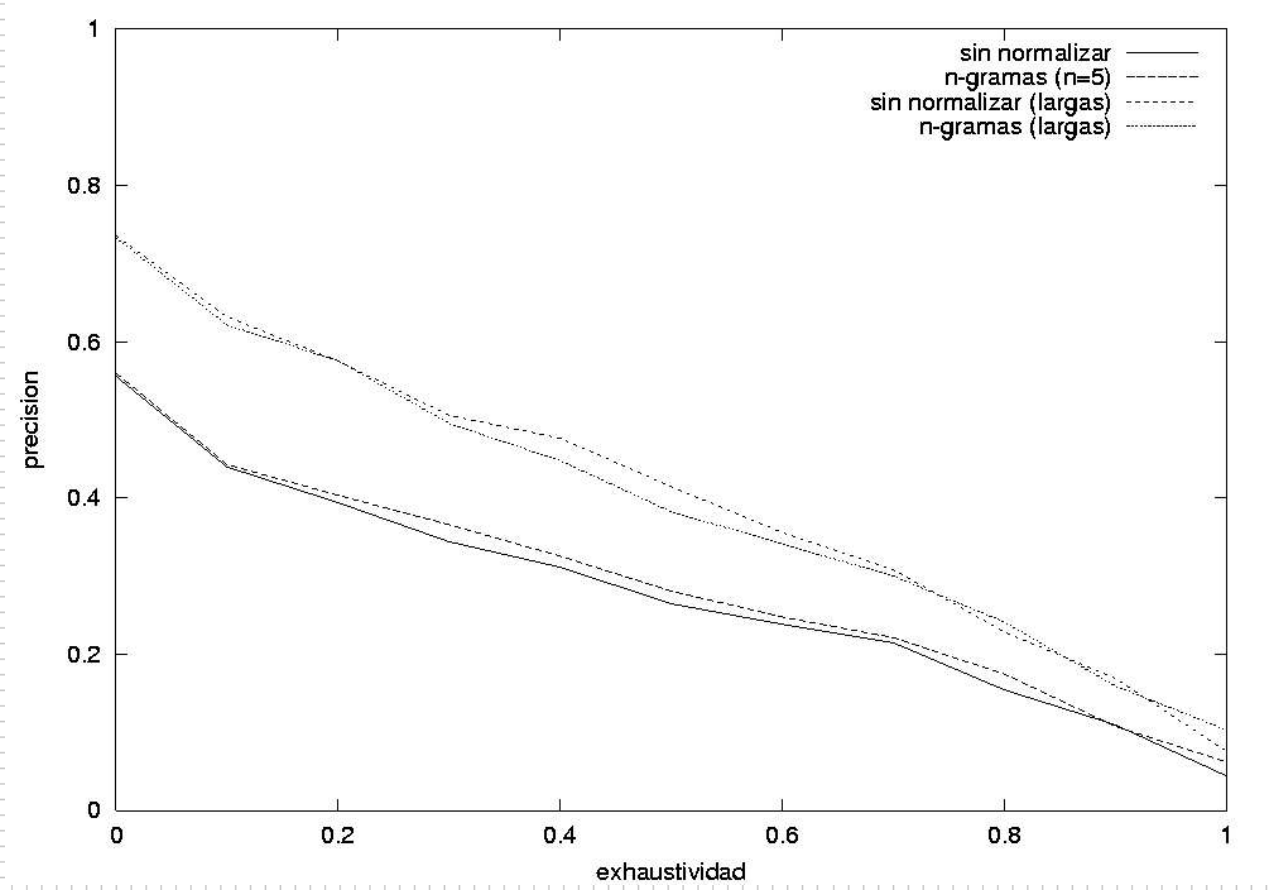


Evaluación. Precisión

Doc.Nº	Simil	Título	
42516	28.00	BANCO HAMBRE PRODUCTOS DE DESECHOS ALIMENTAN A 50.000 PERS	
172743	26.00	UE-AGRICULTURA GREENPEACE DENUNCIA SUBVENCIONES EXPORTACION	
55464	20.00	RFA-ALIMENTOS MAS POTITOS PROCEDENTES DE ESPAÑA CON RASTROS	OK
134812	19.00	MEDIO AMBIENTE AGRICULTURA ECOLOGICA OCUPA 0,1 % TIERRAS CULT	
83220	18.00	EEUU-ALIMENTACION NIÑOS Y JOVENES DE EEUU EXPUESTOS AL CANCER	OK
56832	18.00	RFA-ALIMENTOS PARLAMENTO SE OCUPARA DE POTITOS CONTAMINADOS	OK
49748	18.00	HAITI-EMBARGO DERECHISTAS PROTESTAN POR LLEGADA BUQUE FRA	
121278	17.00	LA CARNE PICADA Y EL POLLO SON LOS ALIMENTOS MAS CONTAMI	OK
133491	16.00	PRECIOS-REACCIONES ECONOMIA ESPERA DESCENSO INFLACION PRO	
46940	16.00	OBESIDAD-TEORIAS DESMIENTEN OBESIDAD SEA CAUSA INGESTION EX	
13245	16.00	CHINA-PESTICIDAS MAS DE 10.000 MUERTOS A CAUSA PESTICIDAS VEN	
56184	15.00	RFA-ALIMENTACION FISCALIA SE OCUPA DE POTITOS CONTAMINADOS	OK
178697	14.00	SUIZA-ALIMENTACION AUMENTA EN UN 50 POR CIENTO CONSUMO ALIM	
175502	14.00	PRESENTADA MADRID FUNDACION "BANCO DE ALIMENTOS DE ESPAÑA"	
126904	14.00	JUBILADOS Y AMAS DE CASA AL FRENTE DE UN BANCO DE ALIMENTOS	
119421	14.00	MEXICO-NIÑOS ALIMENTO DIARIO PARA 64.000 NIÑOS INDIGENAS MEXIC	
108094	14.00	ARGENTINA-MEDIO AMBIENTE DENUNCIAN ENTIERRO CLANDESTINO DE	
85423	14.00	BRASIL-INFLACION GOBIERNO SE PREPARA PARA COMBATIR ESCASEZ	

Precisión = 5 / 18 = 0.27

Evaluación. Gráfico Exhaustividad-Precisión interpolada



(las curvas más alejadas del origen representan mejores resultados)

Indexación con aproximación lingüística

Un breve repaso a los "intentos de mejora"
en el proceso de indexación.

Emilio Rodríguez Vázquez de A.
Universidad de Salamanca
Grupo REINA
http://reina.usal.es

Indexación con aproximación lingüística

2. Motivación
3. Evolución del PLN
4. Indexación morfosintáctica
5. Indexación sintáctica
6. Indexación basada en el sentido de las palabras
7. Conclusiones

1. Motivación

Método básico de indexación de D y Q:

Extracción de palabras ortográficas del texto, normalización (M/m, eliminación de acentos) y supresión de vacías

Es una análisis muy superficial del texto

Implica: sólo recuperaré los documentos que tengan (*algunas o todas*), de mis mismas palabras (teniendo presente, además, que mis palabras se pueden utilizar en contextos que no preveo)

1. Motivación

Q:

Recursos didácticos

[DOC 1]

Este es un *recurso didáctico* de primera magnitud

1. Motivación

Q:

Recursos didácticos

[DOC 2]

Guía de los nuevos *recursos* en la *educación* a distancia

1. Motivación

Q:

Recursos didácticos

[DOC 3]
Se explican aquí los nuevos *medios* audiovisuales para la *enseñanza*

1. Motivación

Q:

Recursos didácticos

[DOC4]

Todos los *recursos* interpuestos, aunque *didácticos*, fueron desestimados

1. Motivación

	Recursos educativos	Recurso educativo	Recursos enseñanza	Medios didácticos
1				
2				
3				
4				
5				
6				
7				
8				
9				
10				

Consulta a Google (9/12/03)

1. Motivación

Otra consulta a Google (9/12/03): “embargo de viviendas” (5 primeros)

1	.. quién da más?” Sin embargo esas viviendas pasan a ser otro foco más de especulación por el abandono de su vigilancia.
2	... ocupadas, que son el conjunto de viviendas principales y secundarias, aumentaron entre 1981 y 1991 en un 27,8%. Sin embargo, las viviendas secundarias son las ...
3	... Existe una variedad de programas y proyectos de HUD que proporcionan este tipo de viviendas, sin embargo estas viviendas a menudo no tienen la estructura...
4	.. En Chile, sin embargo, las viviendas sociales, que representan un alto porcentaje del total de casas que se construyen en el país, no cumplen con este objetivo ...
5	... años. Sin embargo, las viviendas nuevas que no llenen estos requisitos, a partir del año 2005 perderán la exoneración. Las viviendas ...

1. Motivación

Otra consulta a Google (9/12/03): “embargo de viviendas” (5 restantes)

6	... Sin embargo, usted deberá solicitar asistencia para los daños a su finca a ... P. ¿Se puede solicitar préstamos para viviendas secundarios o vacacionales? ...
7	... Sin embargo, llama la atención que las viviendas entregadas entre seis meses y un año más tarde de lo previsto representen nada menos que el 4,11% sobre el ...
8	... Cabe pues esperar que se superen las 500 viviendas por cada 1000 habitantes. Sin embargo, ello no equivale a una mayor disponibilidad de vivienda en general ...
9	... evolución del empleo y los bajos tipos de interés seguirán impulsando la demanda residencial, "que se enfrentará, sin embargo, a viviendas con precios más ...
10	... grupos sin núcleo familiar, etc. Sin embargo, las viviendas se siguen concibiendo mayoritariamente para la primera tipología. ...

1. Motivación

Se ha experimentado (y se experimenta) con otros "métodos de indexación"

Buscando que documentos y preguntas "casen" aún estando en "otras palabras"

Podemos agrupar este conjunto de experimentos (Tzoukermann 97) en:

- Indexación morfosintáctica
- Indexación sintáctica
- Indexación basada en el "sentido de las palabras"

1. Motivación

Para llevarlos a cabo, se han utilizado técnicas y recursos del PLN

También se emplean “técnicas no lingüísticas” (más sencillas)

Indexación con aproximación lingüística

1. Motivación
2. Evolución del PLN
3. Indexación morfosintáctica
4. Indexación sintáctica
5. Indexación basada en el sentido de las palabras
6. Conclusiones

2. Evolución del PLN

La utilización de técnicas lingüísticas ha sido posible gracias al desarrollo, en los 90, de la disciplina conocida como PLN (o LC, o IL, o TL...)

Hitos:

- Irrupción de los modelos probabilísticos y desarrollo y disponibilidad de "Corpus"
- Utilización de técnicas de "estados finitos"
- Técnicas de análisis "robustas" (sobre textos sin restricciones)
- Nuevos ámbitos de aplicación: RI, RT, EI...

Indexación con aproximación lingüística

1. Motivación
2. Evolución del PLN
3. Indexación morfosintáctica
4. Indexación sintáctica
5. Indexación basada en el sentido de las palabras
6. Conclusiones

3. Indexación morfosintáctica

2. Sobre “morfología”
3. Objetivos
4. Indexación morfológica con técnicas no lingüísticas
5. Indexación morfológica con técnicas lingüísticas

3.1. Sobre “morfología”

Morfología: forma de las palabras

Tipos de fenómenos morfológicos:

Flexión:

- comemos(VMPI1P), comen(VMPI3P)
- recurso (NCMS), recursos (NCMP)

Derivación

- educación(N), educar(V), educado(A), educativo(A), educador(N)...

Composición

- moto, sierra -> motosierra
- Podemos incluir la enclisis: saltar, saltarlo, saltárselo

Estos fenómenos no se manifiestan con la misma intensidad en todas las lenguas

3.2. Objetivos

"Uniformización" de variantes morfológicas de las palabras de Q y D:

Juegas, jugábamos, jugaríamos... JUGAR

Ministra, ministro, ministros, ministras MINISTRO

Diseño, diseñador, diseñar,diseñamos... DISEÑ

3.3. I.M. con técnicas no lingüísticas

Eliminación (sin discriminación) normalmente de sufijos (“suffix-stripping”)

Métodos “simples” (eliminación de “s”) o más sofisticados (eliminación de ‘s’,’tion’,’ty’...)

El “inglés” tiene una morfología poco rica

3.3. I.M. con técnicas no lingüísticas

Se han adaptado a diferentes idiomas

Los controles de alcoholemia en las carreteras españolas fueron intensificados por la Guardia Civil y las policías locales en la noche de fin de año en numerosos vías, principalmente en aquellas rutas próximas a locales de diversión, con el fin de evitar que los conductores se hicieran cargo del volante con una copa de más.

l control d alcoholemi en l carreter español fueron intensificad por l guardi civil y l polici local en l noch d fin d año en numeros vi, principalment en aquell rut proxim a local d diversion, con el fin d evitar qu l conductor s hicieran carg del volant con un cop d m.

3.3. I.M. con técnicas no lingüísticas

"Podas" a mayores (algoritmo de Porter):

Antes de "stripping"	Uniformización
university	univer
universe	
organization	organ
organism	
organ	

3.3. I.M. con técnicas no lingüísticas

Los resultados obtenidos, para el inglés, para un simple "s-stemmer" parece que producen ciertas mejoras

Para el "español":

- Efectos positivos, si las preguntas son cortas
- No producen mejora, si las consultas son largas

Ventaja (frente a métodos lingüísticos): menor coste computacional

3.4. I.M. con técnicas lingüísticas

2. Herramientas y recursos necesarios
3. Proceso de indexación
4. Resultados obtenidos

3.4.1. Herramientas y recursos...

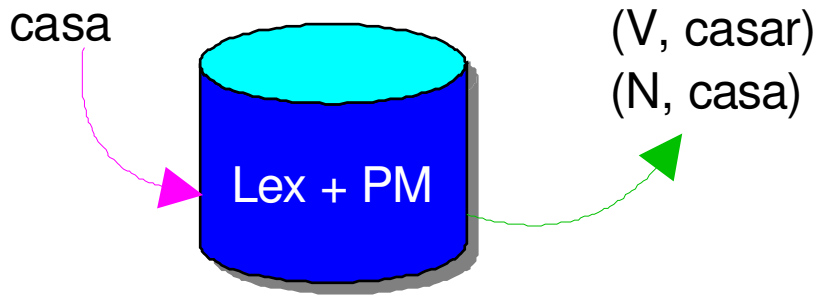
Se necesita:

Un “segmentador” de texto apropiado (*tokenizer*)

Lexicón + procesador morfológico

Desambiguador categorial (o POS-Tagger)

Lexicón y P. Morfológico:



El problema de la morfología computacional puede considerarse “prácticamente” resuelto

3.4.1. Herramientas y recursos...

Proceso de lematización “morfológica” de un texto: “El Príncipe no se casa”

el:D

príncipe:N / príncipe:A

no: R / no: N

se:P / él:P

casa:N / casar:V

Hay que “desambiguar”. Herramienta: Desambiguador Categorial, Part Of Speech (POS) Tagger

3.4.1. Herramientas y recursos...

POS Tagging. Estado del arte:

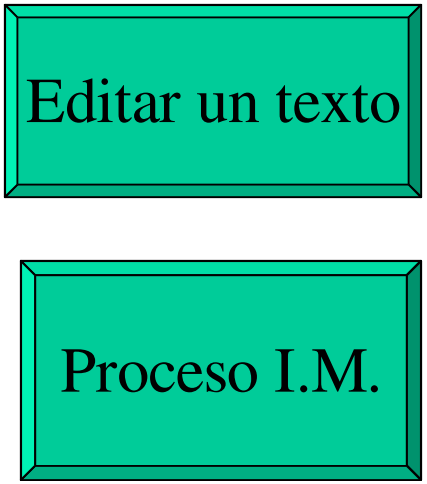
Tipos: basados en reglas, estadísticos e híbridos

La precisión se sitúa entre 95-97%. Algunos "Taggers" por encima

Entrada/Salida de un desambiguador categorial

Entrada	Salida
el (el:D)	el(el:D)
Príncipe (príncipe:N/príncipe:A)	Príncipe(príncipe:N)
no (no:R/no:N)	no (no:R)
se(se:P/él:P)	se(él:P)
casa(casa:N/casar:V)	casa(casar:V)

3.4.2. Proceso de indexación morfológica



En los índices será necesario guardar el par (canónica,categoría):

"El hombre bajo toca el bajo"

No se trata la M. derivativa

3.4.3. Resultados de la investigación

Para el inglés no se consiguen mejoras apreciables respecto de la IM no lingüística. Mayor coste computacional

Incidencia de errores desambiguación: no apreciable

Para otros idiomas: falta comprobación

Ventajas:

- Eliminación coherente de palabras vacías
- Reducción del tamaño del "índice"

Indexación con aproximación lingüística

1. Motivación
2. Evolución del PLN
3. Indexación morfosintáctica
4. Indexación sintáctica
5. Indexación basada en el sentido de las palabras
6. Conclusiones

4. Indexación sintáctica

3. Objetivos
4. Aproximaciones
5. Resultados

4.1. Objetivos

Superar la asunción de independencia de las "palabras" en los métodos de indexación básicos

Reconocer unidades multipalabra, determinadas estructuras sintácticas

P.e.:

- Detección de nombre propios: José María Aznar, George Bush,...
- Conceptos multipalabra: "hot dog", "White House", "Information Retrieval", "bases de datos"...
- Detecta diferencia entre p.e. "college junior" vs "junior college", "Venetian blind" vs "blind Venetian"...

4.2. Aproximaciones

Indexación multipalabra por vía estadística o por vía lingüística

Vía estadística: coocurrencia de términos (se reduce a 2)

Por vía lingüística: llevar a cabo un análisis sintáctico (fundamentalmente SN más o menos complejos)

La vía lingüística requiere añadir al proceso de indexación morfológica una nueva herramienta: Analizador sintáctico superficial (Shallow Parser)

4.3. Resultados

Para colecciones en inglés (grupo Xerox, grupo CLARIT, STRAZALKOWSKI) puede concluirse que:

- Se obtienen ciertas mejoras si las preguntas son largas
- Las técnicas lingüísticas producen mejores resultados que las estadísticas
- Es necesario indexar los “compuestos” también por los “simples”
- ¿Qué peso dar a los compuestos?

Indexación con aproximación lingüística

1. Motivación
2. Evolución del PLN
3. Indexación morfosintáctica
4. Indexación sintáctica
5. Indexación basada en el sentido de las palabras
6. Conclusiones

5. Indexación basada en el “sentido”...

1. Objetivos
2. Indexación con Diccionarios
3. Indexación con Wordnet
4. Desambiguación del sentido de las palabras
5. Resultados

5.1. Objetivos

Fundamentalmente, tratar la polisemia (incluimos homonimia) y/o sinonimia

Como veremos, se busca una indexación a nivel léxico semántico

NO es una Indexación Conceptual

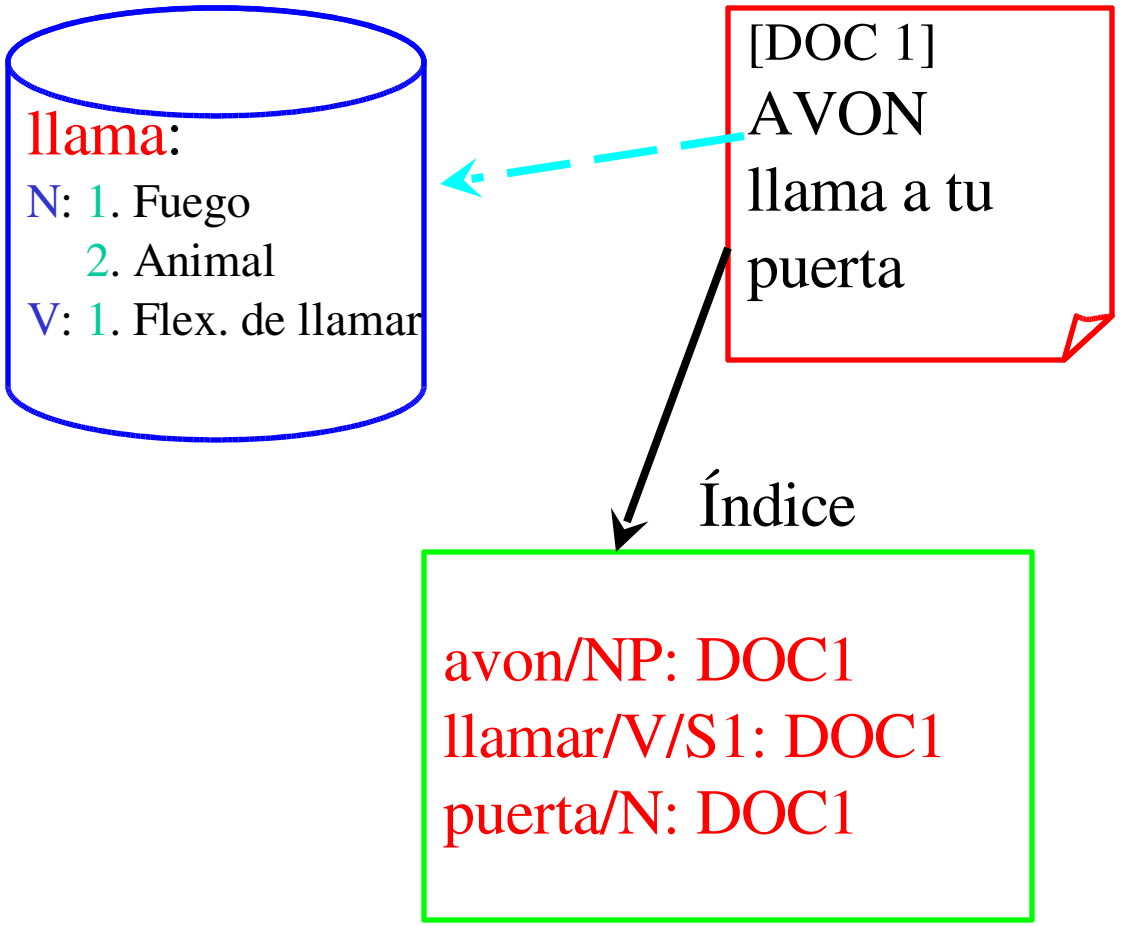
El concejal de urbanismo X de la CA Y ha estado cobrando comisiones de la empresa Z. Dicho concejal se encargó de la recalificación de los terrenos...

X

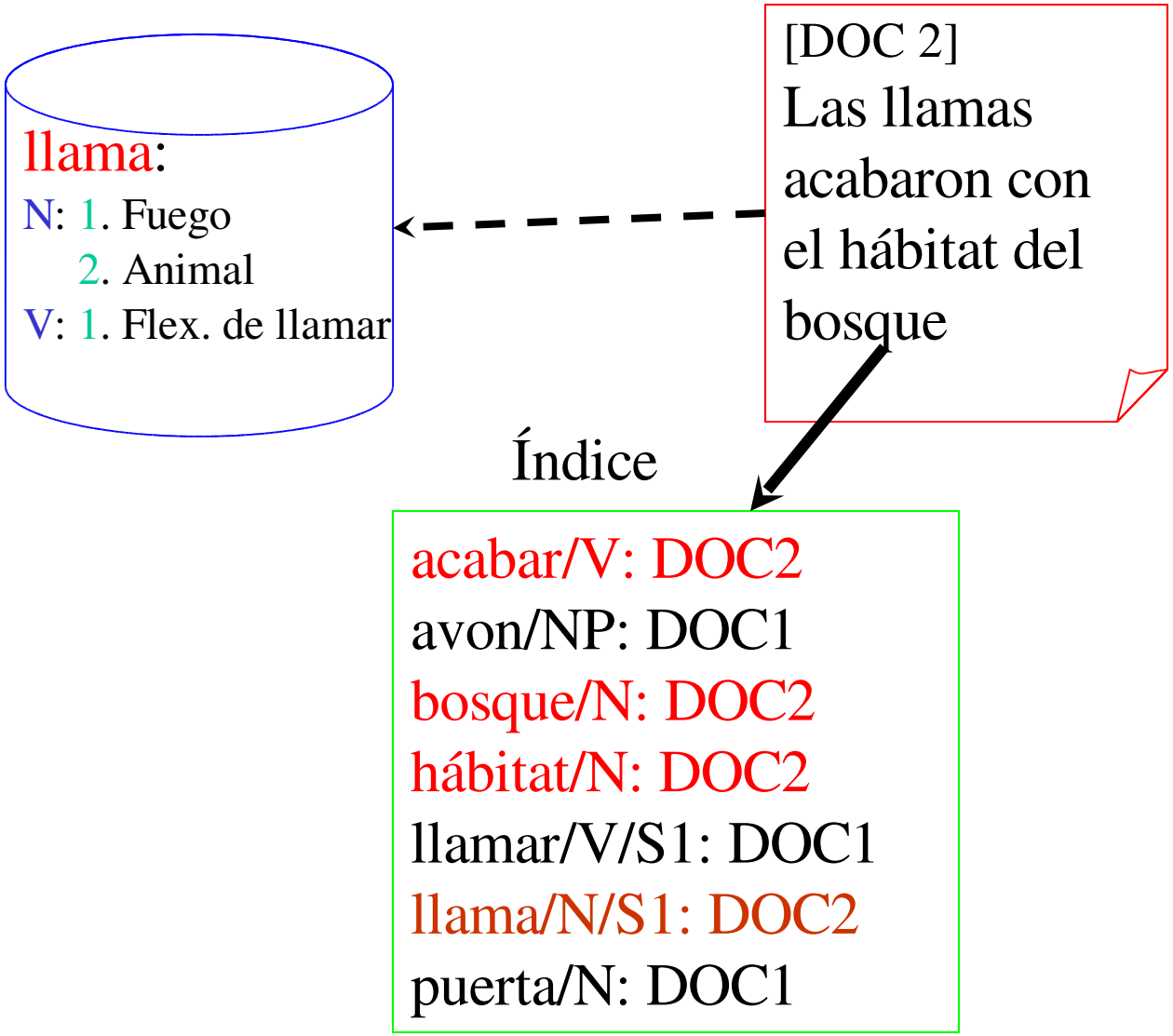
Corrupción política

5.2. Indexación con Diccionarios

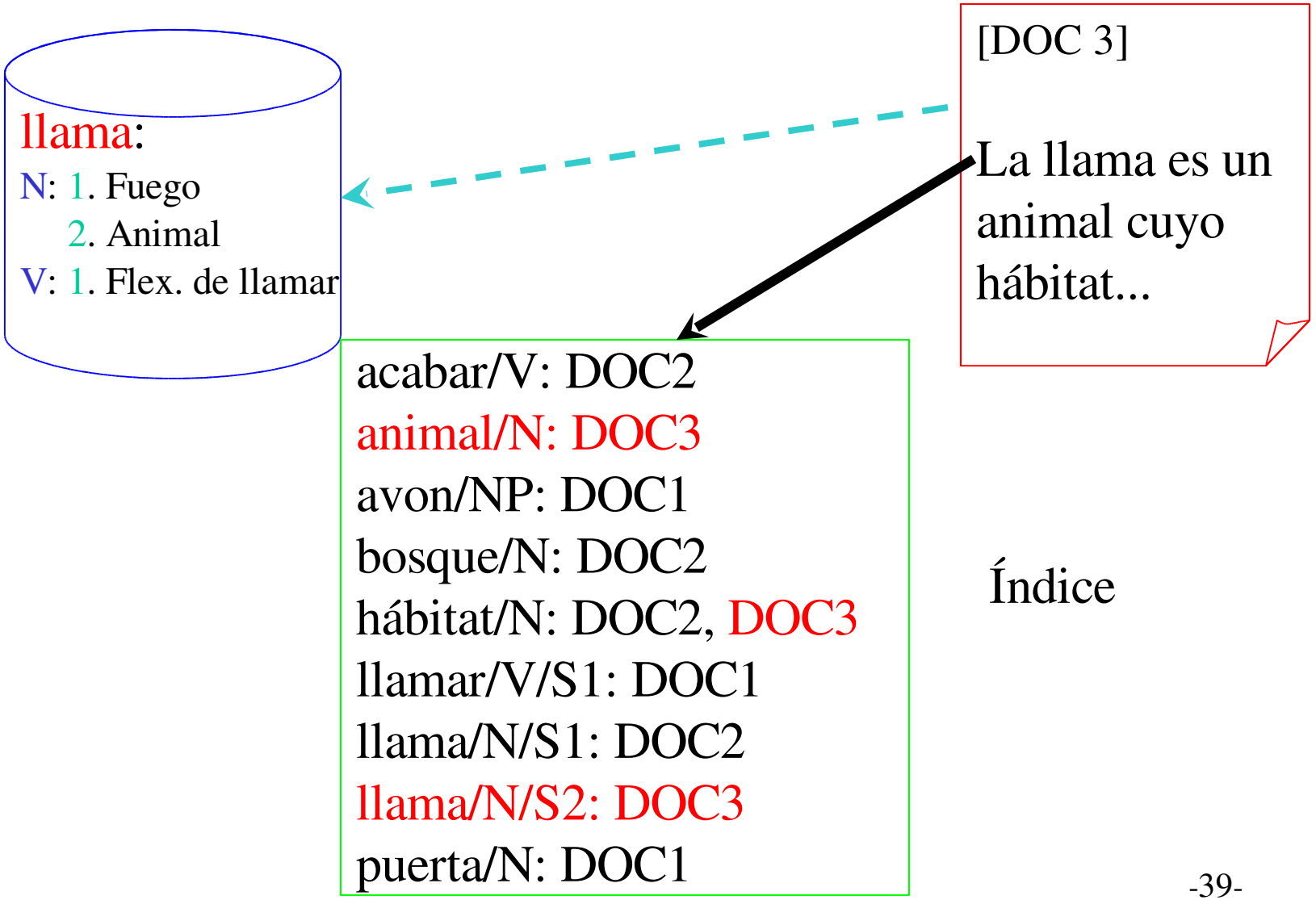
Tratamiento de la polisemia



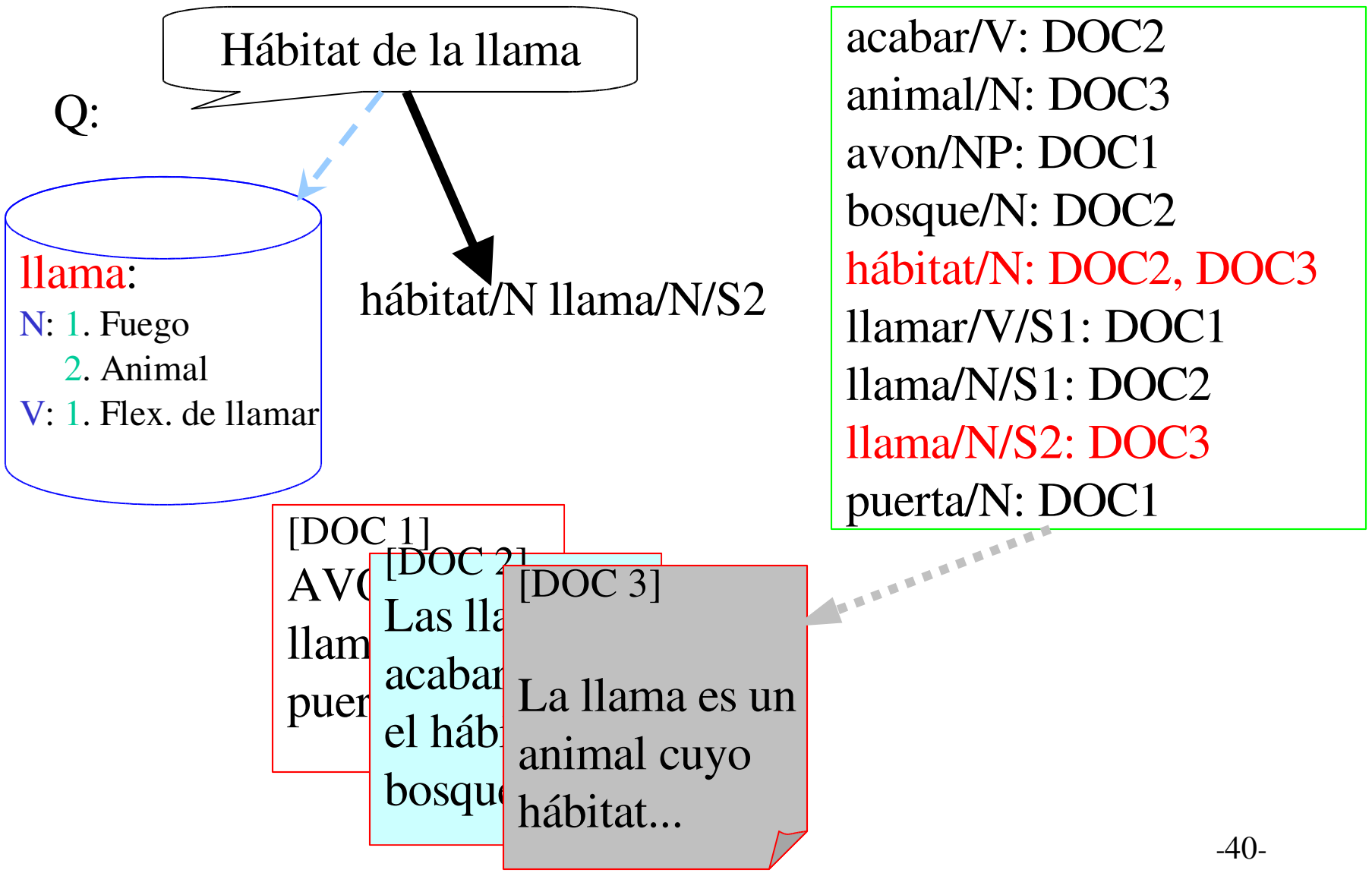
5.2 Indexación con Diccionarios



5.2. Indexación con Diccionarios



5.2. Indexación con Diccionarios



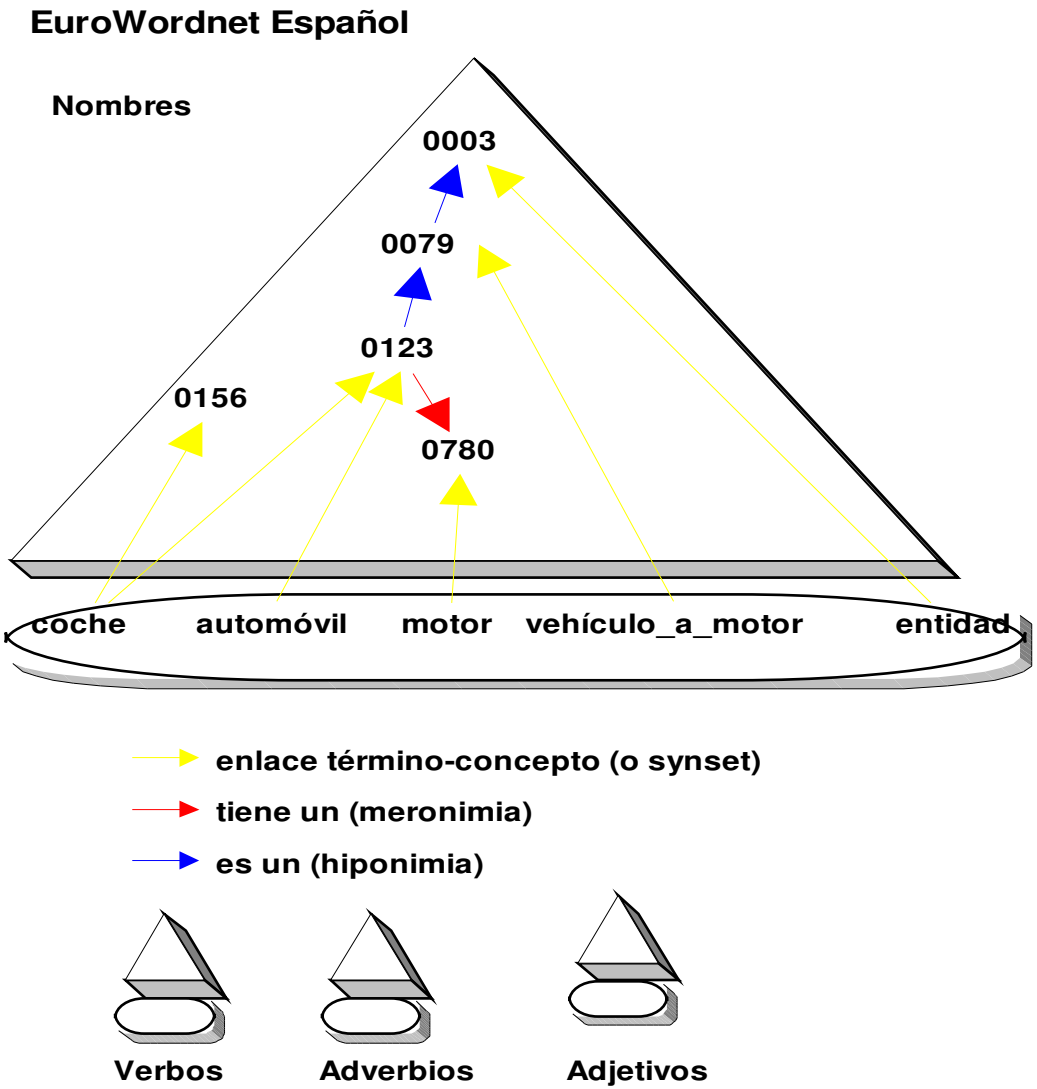
5.3. Indexación con WordNet

Para el tratamiento de la polisemia y la sinonimia es necesario contar con un recurso organizado para tal fin

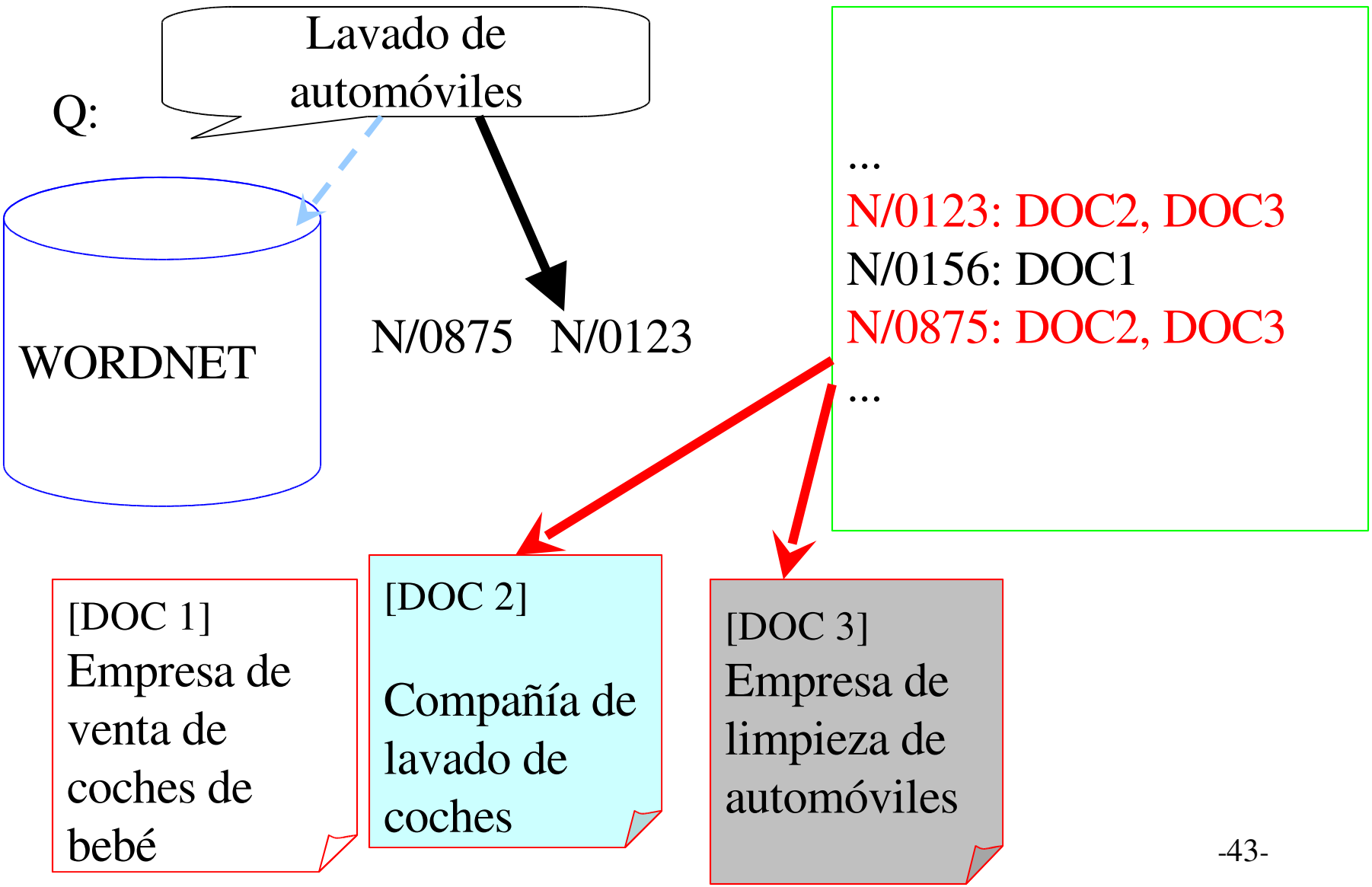
Base de datos WORDNET y EuroWordnet:

Organización básica: el “synset”

5.3. Indexación con Wordnet



5.3. Indexación con Wordnet



5.4. Desambiguación del sentido ...

Word Sense Disambiguation

Objetivo: obtener el sentido correcto de las palabras en un texto, dado el contexto, de forma automática

Tarea muy compleja

Problemas a determinar:

- ¿Qué es "sentido"?
- ¿Qué granularidad es la conveniente?

Aplicaciones: TA, IR, Text Processing...

5.4. Desambiguación del sentido...

Enfoques: enfoque basado en reglas (cualitativo) y cuantitativo

Enfoque cualitativo: sobre dominios restringidos.

Se utilizan recursos léxicos (Diccionarios, Tesaurus y, sobre todo WordNet) y corpus desambiguados (SEMCOR)

Resultados actuales:

En la conferencia SENSEVAL-2 (2001), el mejor sistema consiguió un 69% de precisión

5.5. Resultados

Primero: hay que determinar con qué tasas de error en WSD se produce degradación en RI:

- Difícil de comprobar
- Trabajos de investigación: sitúan el margen de error entre 10% (WSD basada en Diccionarios) y 30% (WSD basada en Wordnet)

No es necesario desambiguar todas las palabras

Parece conveniente trabajar con "granularidad menos fina"

Díficil evaluación

Indexación con aproximación lingüística

1. Motivación
2. Evolución del PLN
3. Indexación morfosintáctica
4. Indexación sintáctica
5. Indexación basada en el sentido de las palabras
6. Conclusiones

6. Conclusiones

Las aportaciones del PLN a RI no han sido las que se esperaban

El frente abierto en la indexación basada en el "sentido de las palabras": no está claro o no está maduro

El tratamiento de la morfología (aunque con técnicas no lingüísticas) parece conveniente

La indexación "sintáctica", en general, no ha producido mejoras

¿Continuar? ¿Reorientar? Hay que mejorar los métodos de RI

Expansión de consultas

- ¿Cómo mejorar los resultados de una primera consulta?
  - El problema
  - Algunas soluciones
  - Resultados experimentales

Ángel F. Zazo (afzazo@usal.es)

Expansión de consultas (i)

- Problema
  - A pesar de tener un buen proceso de indización, frecuentemente los usuarios no encuentran respuestas adecuadas a sus necesidades informativas
  - Inconsistencia de vocabulario: problemas en la asignación de términos a conceptos (sinonimia, polisemia, ...)

    vendo coche usado vs. automóvil de segunda mano
  - El problema es más importante cuanto más corta es la consulta
- Solución:
  - Ampliar términos que mejor definan la necesidad de acuerdo a la colección documental y al modelo de recuperación utilizado

    ⇒ expansión de consultas

Expansión de consultas (ii)

- **Idea**:
  si varios términos están muy relacionados entre sí, cuando un usuario está interesado en uno de ellos, seguramente también lo estará en los otros, y los documentos indizados con términos con una semejanza alta a los utilizados en la consulta también serán relevantes para la necesidad informativa
- La expansión de consultas conlleva:
  - Ampliación de nuevos términos a la consulta original
  - Recálculo de la importancia (el peso en el modelo vectorial) de cada término en la nueva consulta

Expansión de consultas (iii)

- Dificultades
  - Determinar la relación entre términos
    - Gran cantidad de mecanismos
  - Selección de los términos más adecuados para ser añadidos a la consulta original
    - ¿ Todos los relacionados, los que superen un cierto umbral, los que mejor relacionados estén con **toda la consulta** ?
  - La elección del mecanismo de pesado de los términos en la nueva consulta
    - Gran cantidad de de mecanismos
  - No hay que perder de vista el coste computacional para obtener resultados en un tiempo razonable

Expansión de consultas (y iv)

- Clasificación
  - Realimentación de consultas con criterios de relevancia del usuario (user relevance feedback, RF)
  - Expansión automática de consultas (análisis local y global, generalmente se basan en la utilización de clustering):
    - Pseudo realimentación de consultas
    - Análisis del contexto local
    - Utilización de tesauros:
      - **Asociación**
      - **Términos infrecuentes**
      - **Similitud**
      - **Phrase-finder**
      - **Contexto sintáctico**
  - Otros: lematización, etc.

Experientos

- Resultados experimentales sin expandir
- Experimento 1: Lematización
- Experimento 2: Realimentación de consultas
- Experimento 3: Pseudo realimentación de consultas
- Tesauros
  - Experimento 4: Tesauros de asociación global
  - Experimento 5: Tesauros de asociación local
  - Experimento 6: Tesauros de similitud global
  - Experimento 7: Tesauros de similitud local
  - Experimento 8: Tesauros con pasajes de texto

Resultados experimentales sin expandir

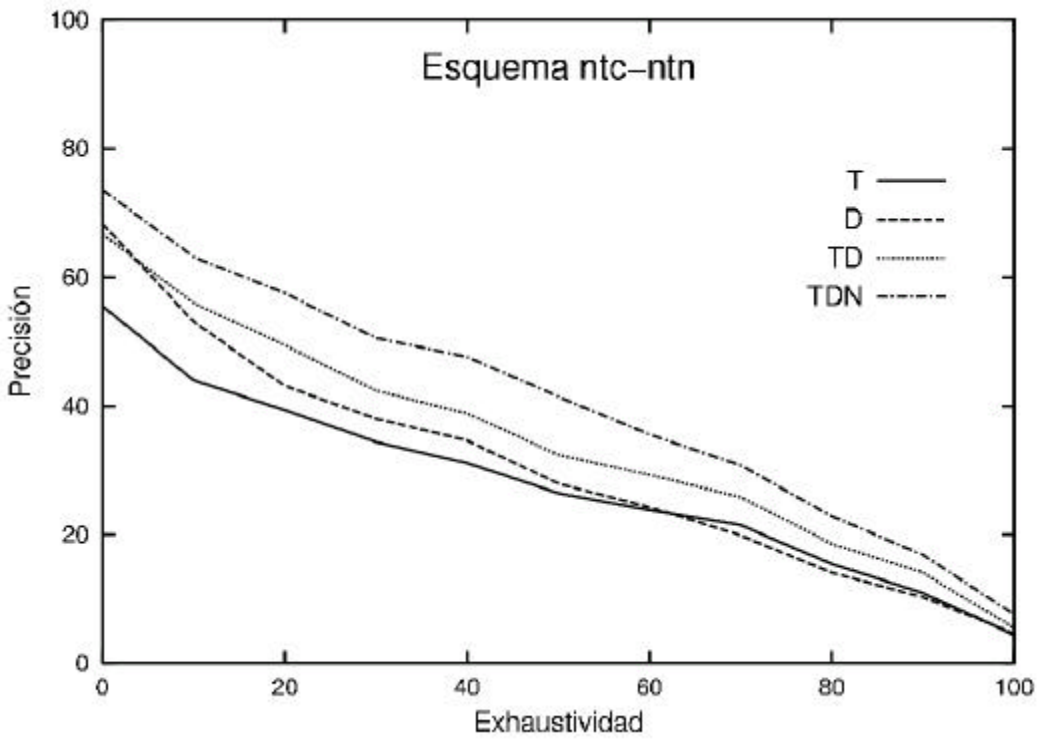
- Experimentos sobre una colección documental:
  - 215.738 documentos (534MB de información) procedentes de la Agencia EFE (noticias de prensa)
  - 50 consultas, con tres campos:

<top> <num> C094 </num>
<ES-title> Retorno de Solzhenitsin </ES-title>
<ES-desc> Encontrar documentos que informen sobre el retorno a Rusia del ganador del premio Nobel de literatura, Solzhenitsin. </ES-desc>
<ES-narr> Los documentos relevantes informarán de las razones y el momento del retorno de Solzhenitsin a Rusia. También pueden mencionar las razones de su emigración a los Estados Unidos. </ES-narr> </top>

	T	D	TD	TDN
Nº de términos en las consultas				
Total	132	421	553	1420
Únicos	123	309	329	653
Nº de términos índice por consulta				
Media	2.64	8.38	8.90	20.46
Desviación	0.79	2.63	2.69	5.77
Máximo	5	15	15	36
Mínimo	1	4	4	10

Resultados experimentales sin expandir

- Diagrama precisión-exhaustividad



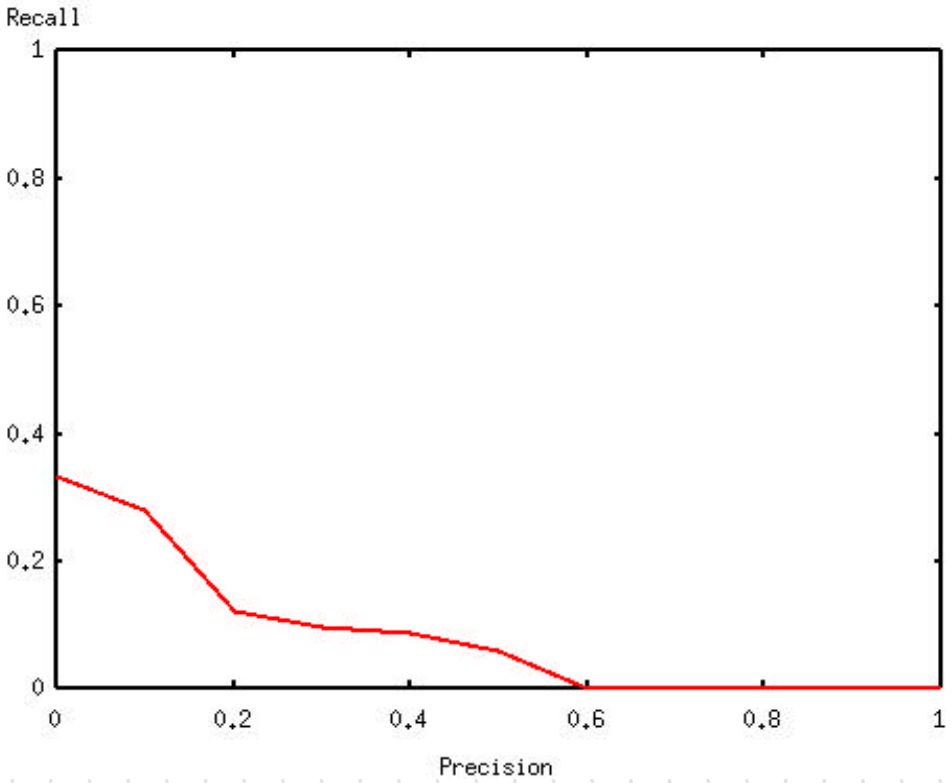
- Precisión a 10 documentos vistos (P@10): 0,3320

Experimento sin expandir. Ejemplo de consulta

Consulta original: "Destrucción de armas nucleares ucranianas"

Pesos: destruccion (4.9160) armas (3.1863) nucleares (4.7264) ucranianas (7.1228)

- Resultados:
  - P. media = 0.0785
  - P@10 = 0.1000



Experimento 1. Lematización

- Cada término de la consulta se expande con aquellos que tienen su mismo lema:
  - Simple (-as, -es, -os, -a, -e, -o)
  - Lematización flexiva (Gómez 2001)
  - Lematización derivativa (Gómez 2001)
- Resultados: mejora en % sobre todas las consultas

	Lem. simple ntc-ntn	Lem. flexiva ntc-ntn	Lem. derivativa ntc-ntn
Prec. media	11.46	4.75	14.87
Prec. P@10	10.84	3.01	13.25

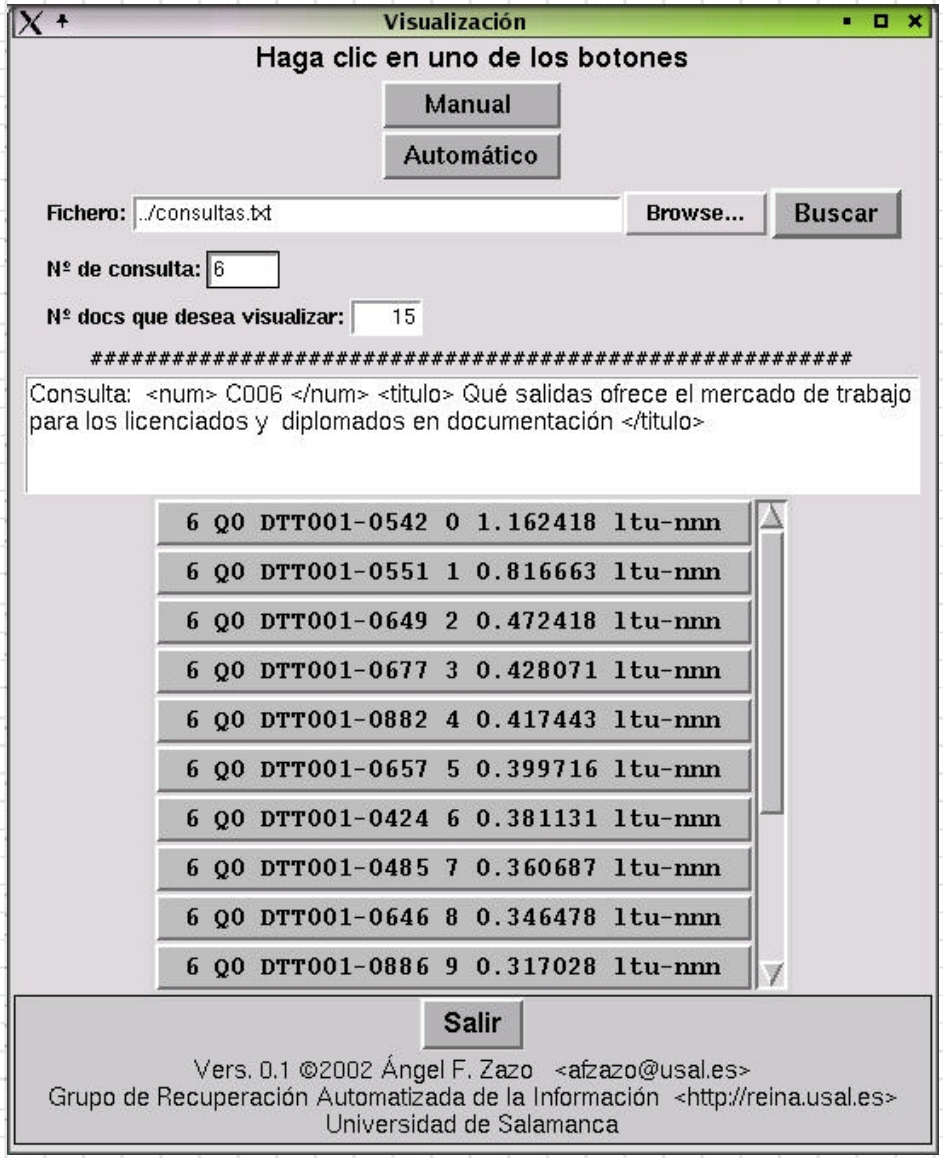
Experimento 2. Realimentación de consultas (i)

- **Usuario**: éste marca documento relevantes y no relevantes
- Rocchio (Rocchio 1979)

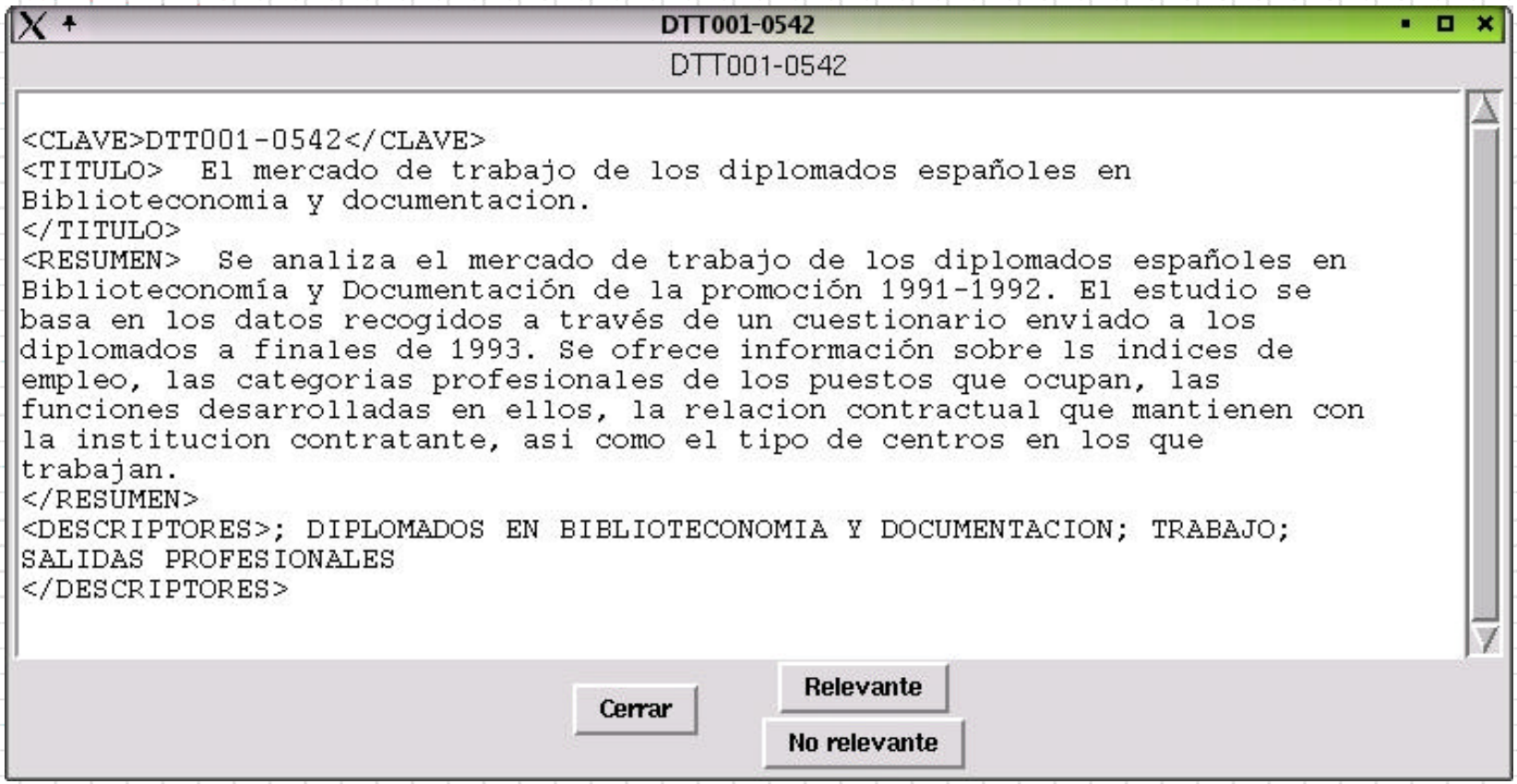
$$\vec{q'} = \alpha\vec{q} + \frac{\beta}{n_{rel}} \sum_{d_j \in rel} \vec{d_j} - \frac{\gamma}{n_{norel}} \sum_{d_j \in norel} \vec{d_j}$$

  - Coeficientes (a, b, g)
  - El número de términos utilizados en la consulta realimentada, en función de su peso
  - Muy buen comportamiento para cualesquiera colección documental, independientemente del esquema de pesado
  - Coeficientes a y b deben ser mayores que g
  - Requiere la presencia del usuario

Experimento 2. Realimentación de consultas (ii)



Experimento 2: Realimentación de consultas (iii)



Experimento 2: Realimentación de consultas (iv)

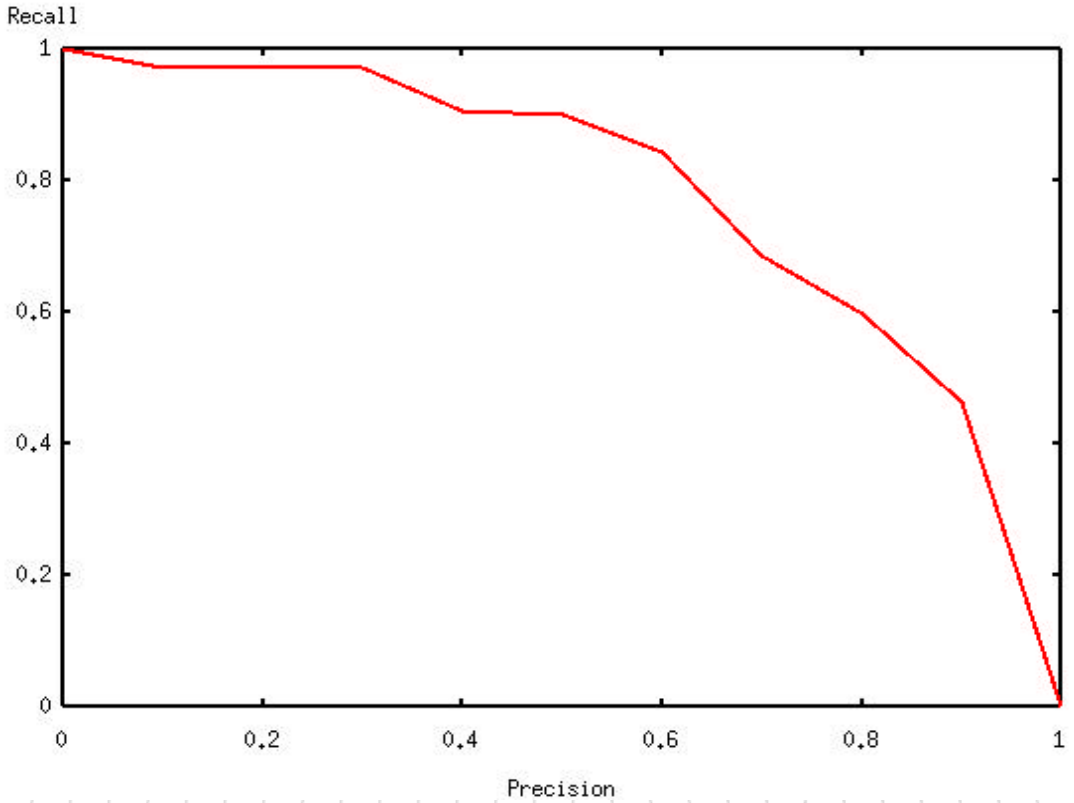
- Consulta original: "Destrucción de armas nucleares ucranianas"
- Documentos relevantes y no relevantes: se han revisado 15 documentos, y el usuario ha marcado 3 relevantes y 12 no relevantes.
- Expansión: primeros términos

Original: ucranianas (7.1228) destruccion (4.9160) nucleares (4.7264) armas (3.1863)

1 - nucleares	64.5947	8 - moscu	22.4899	15 - cabezas	15.3937
2 - ucrania	47.4910	9 - clinton	19.7605	16 - nunn	14.5028
3 - ucranianas	38.5816	10 - convoy	18.2745	17 - arsenales	13.8613
4 - rusia	33.5297	11 - cargas	18.1013	18 - eeuu	12.4935
5 - kravchuk	25.3561	12 - armas	18.0559	19 - mikailov	12.2818
6 - gauldin	24.5636	13 - energia	17.9704	20 - ruso	12.0008
7 - desmantelamiento	24.1274	14 - destruccion	17.6157	-	

Experimento 2: Realimentación de consultas (y v)

- Resultados para esta pregunta:
- P.media = 0.7830
- P@10 = 1.0



Experimento 3: Pseudo realimentación de consultas (i)

- Mecanismo **automático** de expansión
- Rocchio: combinaciones (a, b, $g = 0$)
- Se consideran relevantes los 20, 15, 10 ó 5 primeros documentos recuperados
- Resultados:
  - Se mejoran los resultados medios si a es mayor que b (o al menos no se empeoran)
    - primeros 5 documentos, entre 30 y 50 términos
  - Se realiza en tiempo real
  - Depende de cuán buenos hayan sido los resultados de la primera consulta (mecanismo de pesado)
  - Algunas consultas empeoran (cortas/ambiguas)

Experimento 3: Pseudo realimentación de consultas (ii)

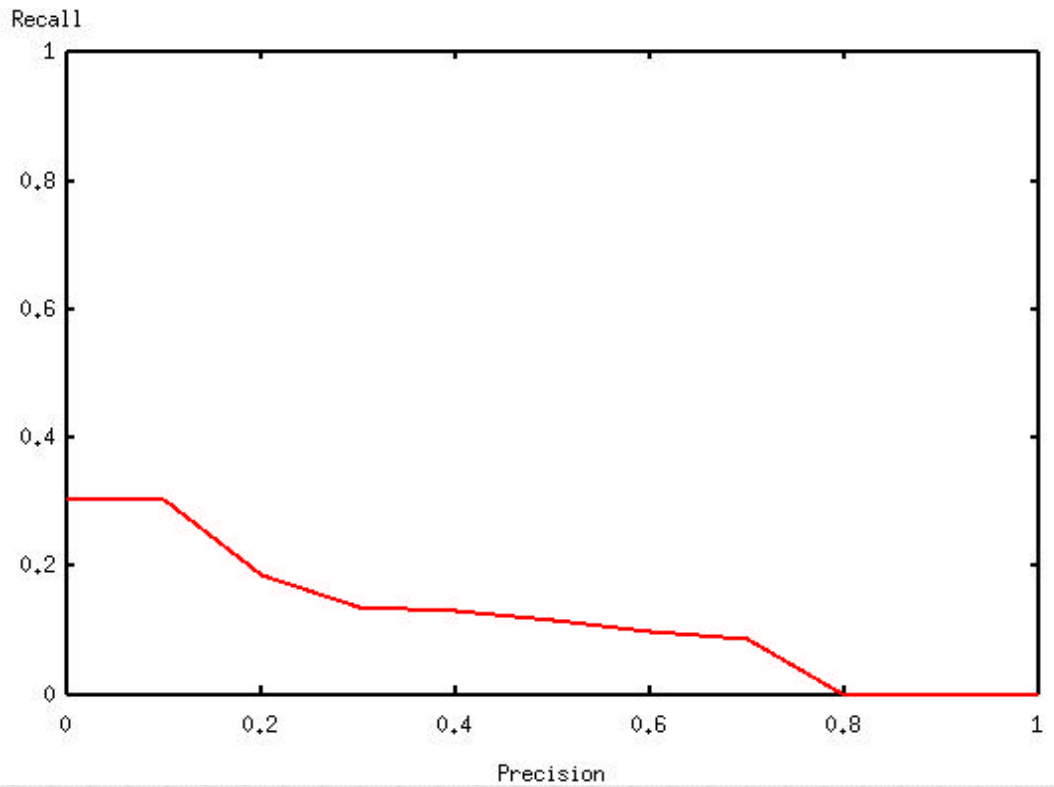
- Consideramos relevantes los 5 primeros documentos recuperados, y consideramos los primeros 40 términos en la consulta realimentada
- $a = 1, b = 0.1$

Original: ucranianas (7.1228) destruccion (4.9160) nucleares (4.7264) armas (3.1863)

1 - nucleares	7.4677	14 - bacteriologicas	0.7607	27 - armamento	0.3981
2 - ucranianas	7.2652	15 - belem	0.6874	28 - ucrania	0.3896
3 - destruccion	6.2924	16 - proscripcion	0.6835	29 - estados	0.3845
4 - armas	5.7991	17 - asamblea	0.5998	30 - kravchuk	0.3803
5 - proliferacion	1.6213	18 - rusia	0.5325	31 - prohibicion	0.3787
6 - tratado	1.2024	19 - acuerdos	0.4633	32 - proscribe	0.3756
7 - oea	1.0407	20 - compremeten	0.4473	33 - quimicas	0.3739
8 - velayati	1.0118	21 - nuclear	0.4336	34 - ensayos	0.3650
9 - masiva	0.8973	22 - pruebas	0.4330	35 - uso	0.3534
10 - desarme	0.8632	23 - continente	0.4190	36 - adheridos	0.3514
11 - resolucion	0.8160	24 - ginebra	0.4155	37 - clinton	0.3487
12 - tlatelolco	0.8077	25 - asfixiantes	0.4033	38 - brasil	0.3485
13 - paises	0.7964	26 - do	0.3991	39 - completo	0.3478

Experimento 3: Pseudo realimentación de consultas (y iii)

- Resultados:
- P. media = 0.1137
- P@10 = 0.2000



Tesauros (i)

- Tesauro: matriz que mide relaciones entre términos

- Nuestros experimentos
  - Tesauros de asociación: valores de coocurrencia
  - Tesauros de similitud: transposición de la matriz documentos-términos

- Consideraciones
  - Tesauro global vs. tesauro local

Tesauros (y ii)

- Expansión. Importante
  - Cómo se contruye el tesauro
  - Selección de términos añadidos
    - Todos los relacionados ¿?
    - Aplicar un umbral ¿?
    - Utilizar los términos mejor relacionados con toda la consulta

$$\mathrm{sim}(q,t) = \mathrm{sim}(\sum_{t_i \in q} w_{iq} t_i, t) = \sum_{t_i \in q} w_{iq} \cdot \mathrm{REL}(t_i, t)$$

  - Pesado de términos añadidos
    - Coeficiente de reducción en función de la consulta original

Unidad	Longitud	Suma	Módulo
1	$\frac{1}{n_q}$	$\frac{1}{\sum_{t_i \in q} w_{iq}}$	$\frac{1}{\lvert q \rvert}$

Experimento 4: Tesauros de asociación global (i)

- Medidas de coocurrencias:

$$\text{Tanimoto}(t_i, t_k) = \frac{n_{ik}}{n_i + n_k - n_{ik}}$$

$$\text{Coseno}(t_i, t_k) = \frac{n_{ik}}{\sqrt{n_i \cdot n_k}}$$

$$\text{Dice}(t_i, t_k) = \frac{2 \cdot n_{ik}}{n_i + n_k}$$

- Resultados:
  - Forma muy simple de obtener relaciones entre términos
  - Buena técnica de expansión de consultas, pese a estar muy denostada. Creemos que se debe a la forma de realizar la expansión.
  - Mejor coeficientes de reducción: 'Suma'
  - Número de términos añadidos a la consulta original
    - Entre 50 y 75 términos

Experimento 4: Tesauros de asociación global (ii)

- Ejemplo: entrada "espacial"

Tanimoto		Coseno		Dice	
espacial	1.0000	espacial	1.0000	espacial	1.0000
nasa	0.2788	astronautas	0.4860	nasa	0.4361
astronautas	0.2670	nasa	0.4755	astronautas	0.4215
espaciales	0.2583	espaciales	0.4472	espaciales	0.4105
orbita	0.2336	cañaveral	0.4002	orbita	0.3788
transbordador	0.2026	astronauta	0.3935	transbordador	0.3369
astronauta	0.1846	orbita	0.3838	astronauta	0.3117
cañaveral	0.1809	orbital	0.3609	cañaveral	0.3063
orbital	0.1540	transbordador	0.3443	orbital	0.2669
nave	0.1339	shuttle	0.3250	nave	0.2361
shuttle	0.1310	discovery	0.3088	shuttle	0.2317
aeronautica	0.1250	soyuz	0.2461	aeronautica	0.2222
cohete	0.1196	transbordadores	0.2450	cohete	0.2136
satelites	0.1153	nave	0.2377	satelites	0.2067
transbordadores	0.1152	hubble	0.2372	transbordadores	0.2066
discovery	0.1121	cosmonautas	0.2335	discovery	0.2016
satelite	0.1106	cohete	0.2284	satelite	0.1991
experimentos	0.1049	cosmonauta	0.2283	experimentos	0.1899
jirl	0.0970	espacio	0.2242	jirl	0.1768
kennedy	0.0928	aeronautica	0.2225	kennedy	0.1698
endeavour	0.0874	atlantis	0.2221	endeavour	0.1607

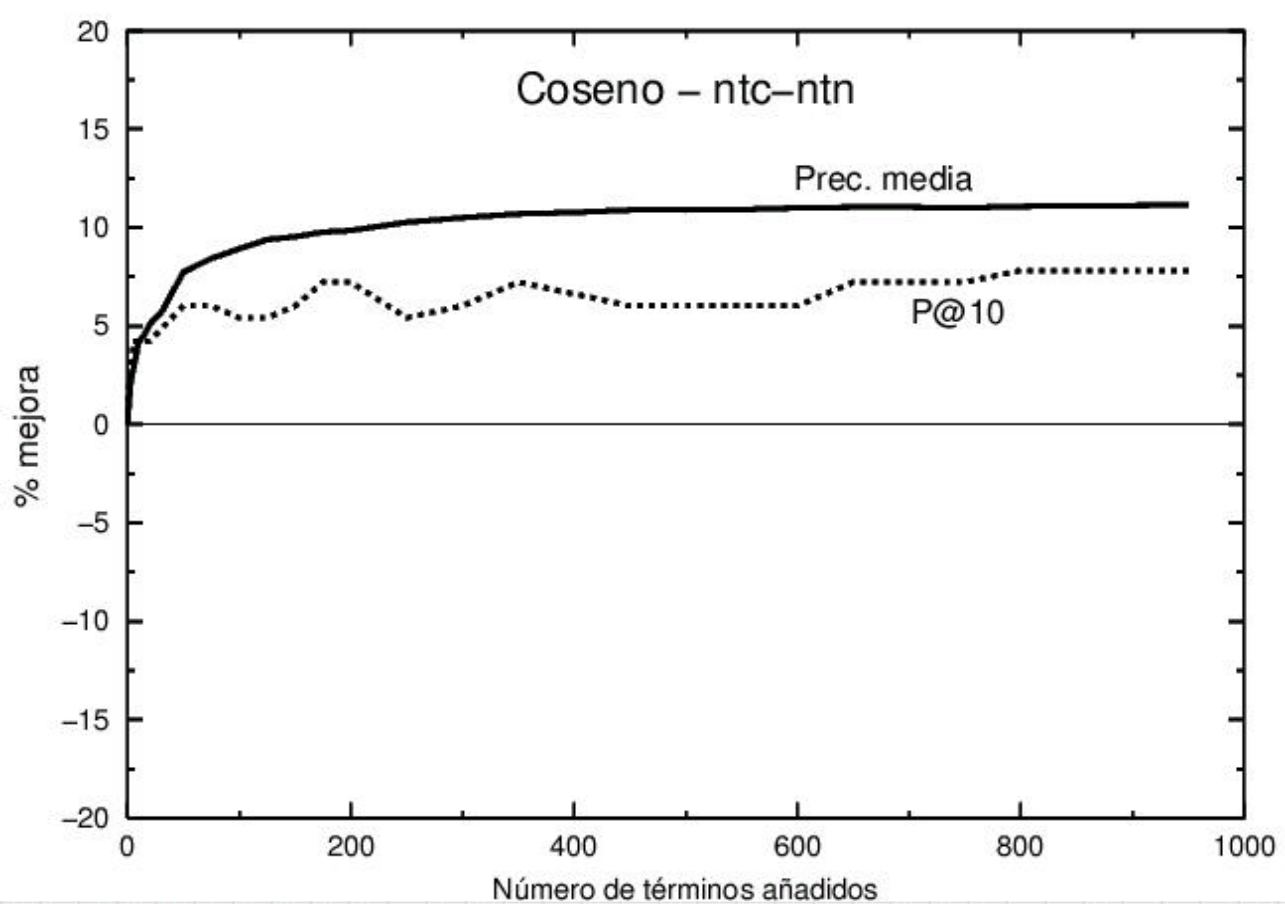
Experimento 4: Tesauros de asociación global (iii)

- Consulta: "Destrucción de armas nucleares ucranianas"
- Primeros términos relacionados (función coseno)

Original: ucranianas (7.1228) destruccion (4.9160) nucleares (4.7264) armas (3.1863)					
1 - nuclear	0.1962	9 - atomicas	0.1287	17 - sebastopol	0.0986
2 - ucrania	0.1519	10 - ucraniano	0.1285	18 - peninsula	0.0960
3 - kiev	0.1489	11 - proliferacion	0.1283	19 - crimeano	0.0935
4 - atomica	0.1485	12 - oiea	0.1207	20 - misiles	0.0928
5 - kravchuk	0.1451	13 - ucranianos	0.1192	21 - moscu	0.0909
6 - leonid	0.1445	14 - pyongyang	0.1178	22 - tnp	0.0899
7 - ucraniana	0.1319	15 - plutonio	0.1084	23 - corea	0.0891
8 - crimea	0.1313	16 - rusia	0.1009	24 - meshkov	0.0886

Experimento 4: Tesauros de asociación global (y iv)

- Resultados en % sobre una colección de 50 preguntas, en función del número de términos añadidos



Experimento 5: Tesauros de asociación local (i)

- Se toman los primeros documentos para construir el tesauro: ¿cuántos?
  - EFE: 1000, 100, 50 y 10 documentos
- Resultados:
  - Mejor cuanto más local es el tesauro (10 documentos)
  - Mejor coeficiente: 'Suma'
  - Número términos añadidos: entre 25 y 50
  - Mejores resultados que los tesauros globales, pero se debe realizar en tiempo real

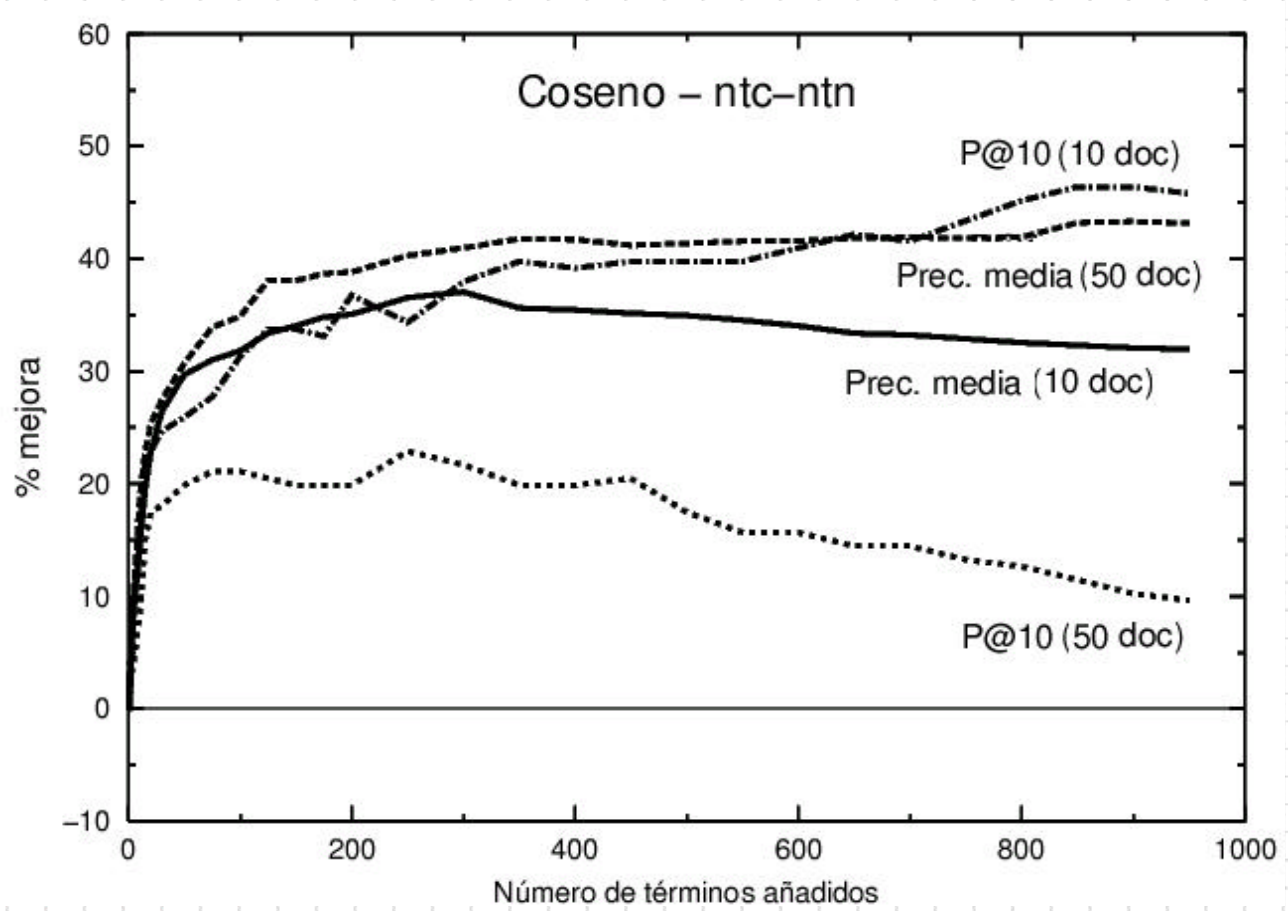
Experimento 5: Tesauros de asociación local (ii)

- Consulta: "Destrucción de armas nucleares ucranianas"

Tesauro de asociación local: 100 documentos					
1 - nuclear	0.5215	8 - paises	0.4539	15 - kravchuk	0.4158
2 - rusia	0.4892	9 - armas	0.4475	16 - ruso	0.4151
3 - seguridad	0.4859	10 - desarme	0.4461	17 - ucraniano	0.4017
4 - moscu	0.4786	11 - presidente	0.4440	18 - eeuu	0.3939
5 - estados	0.4723	12 - leonid	0.4268	19 - boris	0.3922
6 - ucrania	0.4639	13 - unidos	0.4237	20 - yeltsin	0.3922
7 - tratado	0.4618	14 - pais	0.4196	21 - guerra	0.3856
Tesauro de asociación local: 50 documentos					
1 - nuclear	0.5579	8 - presidente	0.4665	15 - ruso	0.4388
2 - rusia	0.5063	9 - unidos	0.4602	16 - armas	0.4384
3 - seguridad	0.5059	10 - desarme	0.4482	17 - viernes	0.4302
4 - moscu	0.5050	11 - eeuu	0.4454	18 - kravchuk	0.4068
5 - estados	0.4965	12 - paises	0.4415	19 - leonid	0.4068
6 - ucrania	0.4786	13 - bill	0.4388	20 - ucraniano	0.4058
7 - tratado	0.4764	14 - clinton	0.4388	21 - pais	0.3975
Tesauro de asociación local: 10 documentos					
1 - seguridad	0.6770	8 - 19	0.5361	15 - militares	0.5170
2 - estados	0.6329	9 - intencion	0.5361	16 - rusa	0.5170
3 - paises	0.6329	10 - presidente	0.5361	17 - rusia	0.5170
4 - nuclear	0.5919	11 - guerra	0.5346	18 - miercoles	0.5161
5 - acuerdos	0.5847	12 - 01	0.5344	19 - internacional	0.4935
6 - ginebra	0.5847	13 - 09	0.5344	20 - proliferacion	0.4935
7 - unidos	0.5847	14 - armas	0.5232	21 - tratado	0.4935

Experimento 5: Tesauros de asociación local (y iii)

- Resultados en % sobre una colección de 50 preguntas



Experimentos 6 y 7: Tesauros de similitud global y local

- Tesauros de similitud: inversión del espacio vectorial de documentos (EVD) ⇒ espacio vectorial de términos (EVT)
- Construcción:
  - Transposición simple: multitud de esquemas de pesado:
  - Aplicación del factor itf (inverse term frecuency)
- Resultados:
  - Coste computacional elevadísimo
  - Ligeramente es mejor la aplicación del factor itf que la transposición simple
  - Se obtienen prácticamente los mismos resultados que con los tesauros de asociación (global y local, respectivamente)

Experimento 8: Tesauros de asociación global con pasajes de texto

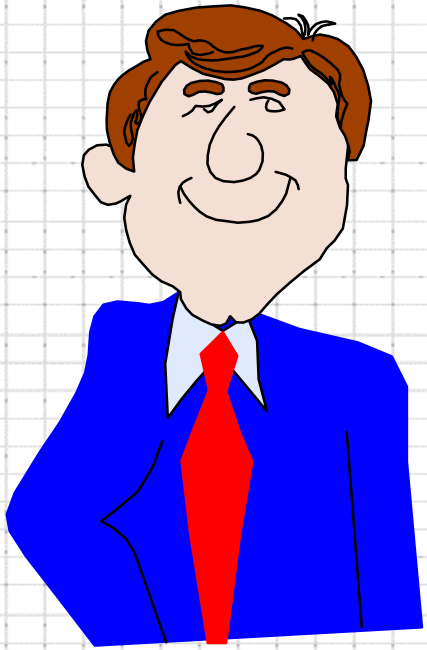
- División del documento en pasajes de texto: ¿tamaño?
  - Ventanas de 300, 200, 100 y 50 palabras
  - Pasajes formados por párrafos y frases
- Consideración:
  - Coste computacional aumenta (división en pasajes)
  - Documentos monotemáticos
- Resultados:
  - Prácticamente mismos resultados que con el documento completo, peores cuanto más pequeño es el pasaje de texto, debido a que los documentos son **monotemáticos**.

Resumen de resultados (i)

Técnica de expansión	Colección	% de mejora	Comentarios
Lematización simple			
	EFE	Prec.: 11 - 12% P@10: 7 - 10%	
Pseudo Realimentación de Consultas			
	EFE	Prec.: 10- 16% P@10: 16 - 8%	Primeros 5 documentos $\alpha > \beta$ entre 30 y 50 términos en la consulta
Tesauros de asociación global (ídem similitud)			
	EFE	Prec.: 5 - 10% P@10: 2 - 5%	Entre 50 y 75 términos en la consulta
Tesauros de asociación local (ídem similitud)			
	EFE	Prec.: 20 - 30% P@10: 10 - 20%	Primeros 10 documentos Entre 25 y 50 términos en la consulta

Resumen de resultados (y ii)

- Mejor es la expansión cuanto peor es la recuperación inicial (= esquema de pesado)
- Es mejor un buen esquema de pesado (modelo de recuperación) que un buen método de expansión
  - Pero obtener un buen esquema de pesado es muy difícil en colecciones que no se han puesto en servicio o colecciones cambiantes (Web)
- La expansión no mejora todas las consultas por igual:
  - Consultas cortas
  - Diferente ambigüedad semántica
  - Diferente calidad en la expansión utilizando técnicas locales en función del esquema de pesado (modelo de recuperación)
- La aplicación de técnicas no lingüísticas para la expansión son muy aceptables, con un gasto computacional muy por debajo de las que son necesarias con técnicas lingüísticas



Y ahora, ¿tiene alguna pregunta?

Ángel F. Zazo (afzazo@usal.es)

Métodos de Recuperación de Información en el Web.

Métodos de RI en el web

- Las técnicas de RI empleadas en el web proceden de los SRI tradicionales.
  Por ello surgen grandes problemas pues el entorno de trabajo no es el mismo y las características de los datos almacenados difieren considerablemente.

- Hay nuevos problemas como el spamming o el enorme tamaño de los índices, haciendo difícil su adecuada gestión con los modelos tradicionales.

Métodos de RI en el web

Hay básicamente tres formas de buscar información:

- Emplear motores de búsqueda.
- Empleo de directorios.
- Buscar explotando la estructura hipertextual.

Métodos de RI en el web

Motores de búsqueda:

- Indexan una porción de los documentos residentes en la globalidad del web.
- Permiten la localización de la información a través de la formulación de una pregunta.

Métodos de RI en el web

Hay básicamente tres formas de buscar información:

- Emplear motores de búsqueda.
- Empleo de directorios.
- Buscar explotando la estructura hipertextual.

Métodos de RI en el web

Directorios:

- Clasifican los documentos web por materias.
- Podemos navegar por las secciones o buscar en sus índices.

Métodos de RI en el web

Características de motores y directorios:

	Descubrimiento de recursos	Representación del contenido	Representación de la consulta	Presentación de los resultados
Motores de búsqueda	Automática por robots	Indización automática	Explícita (palabras clave, operadores)	Páginas creadas dinámicamente en cada consulta. Exhaustivos y poco precisos
Directorios	Lo realizan las personas	Clasificación manual	Implícita (navegación por categorías)	Páginas creadas antes de la consulta. Poco exhaustivos, muy precisos.

Métodos de RI en el web (II)

- Hay básicamente tres formas de buscar información:
  - Emplear motores de búsqueda.
  - Empleo de directorios.
  - Buscar explotando la estructura hipertextual.

Métodos de RI en el web

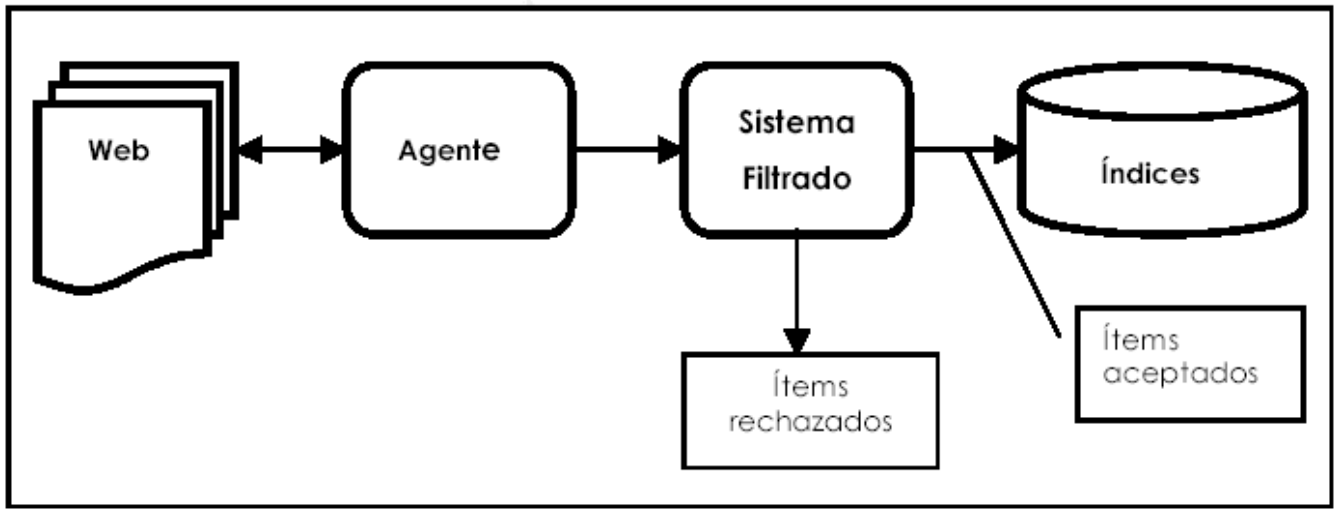
Explotación de la estructura hipertextual:

- Es un mecanismo poco utilizado.
- Son mecanismos difíciles de implantar para grandes cantidades de información .

Métodos de RI en el web

Algunos autores añaden a estos tres sistemas comentados los siguientes:

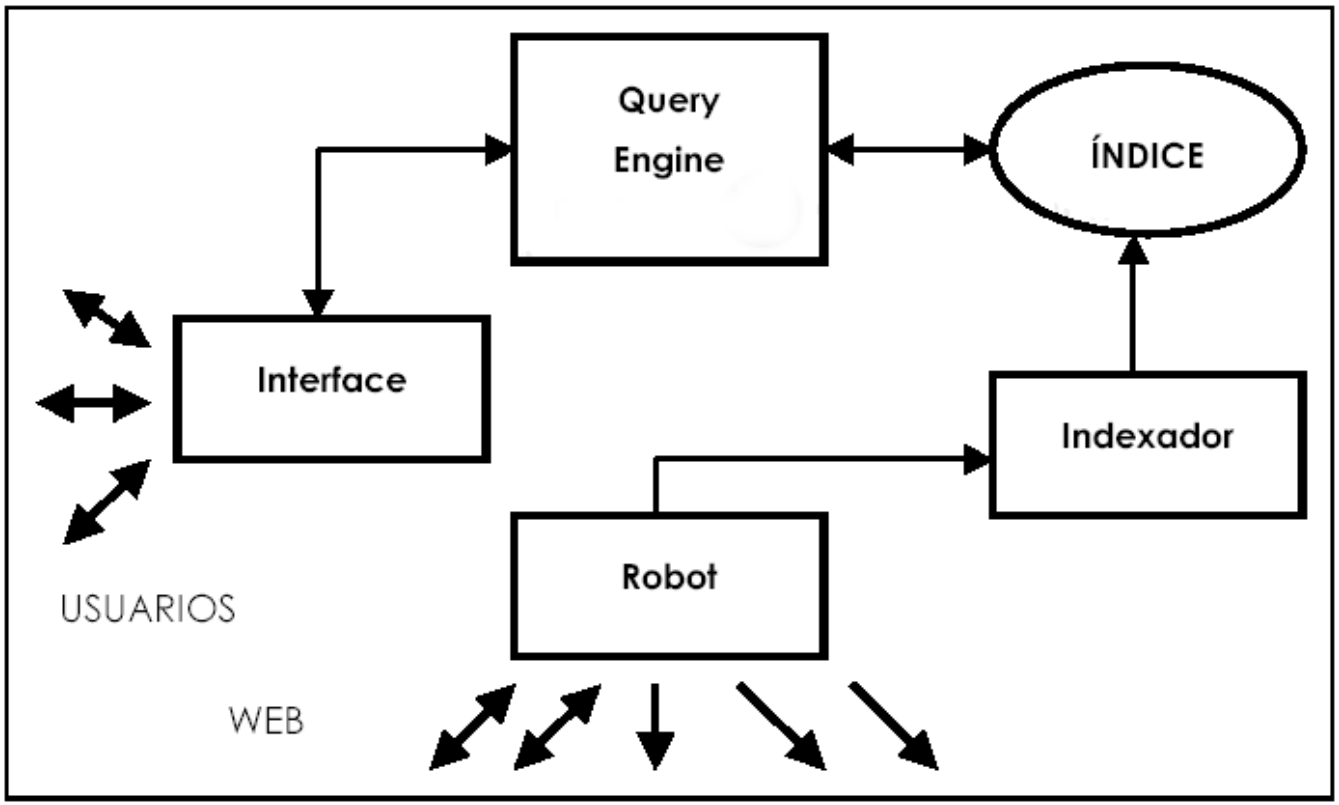
- ✓ Metabuscadores
- ✓ Filtrado de información



Métodos de RI en el web

Funcionamiento de los motores de búsqueda

Arquitectura robot-indexador



Métodos de RI en el web

Funcionamiento de los motores de búsqueda

Los robots son programas que de forma automática permiten rastrear el web. Inician el rastreo a partir de una dirección URL y se siguen los enlaces contenidos en esa URL.

Otras modalidades son:

- ✓ **Knowbots:** programados para localizar referencias hipertexto dirigidas hacia un documento. Permiten evaluar el impacto de las diferentes aportaciones de áreas del conocimiento
- ✓ **Wanderers (vagabundos):** Encargados de realizar estadísticas
- ✓ **Worms (gusanos):** Encargados de la duplicación de directorios ftp
- ✓ **WebAnts (hormigas):** Conjunto de robots, alejados físicamente, que cooperan

Métodos de RI en el web

Funcionamiento de los motores de búsqueda

El índice es el corazón del motor de búsqueda. Normalmente consiste en una lista de palabras asociadas a sus correspondientes documentos. Normalmente se emplea un fichero inverso del tipo:

Document	Text
1	Pease porridge hot, pease porridge cold,
2	Pease porridge in the pot,
3	Nine days old.
4	Some like it hot, some like it cold,
5	Some like it in the pot,
6	Nine days old.

(a) Example text; each line is one document

Number	Term	Text
1	cold	1,4
2	days	3,6
3	hot	1,4
4	in	2,5
5	it	4,5
6	like	4,5
7	nine	3,6
8	old	3,6
9	pease	1,2
10	porridge	1,2
11	pot	2,5
12	some	4,5
13	the	2,5

(b) Inverted file for text of (a)

Métodos de RI en el web

El fichero inverso se convierte en una enorme estructura de datos con problemas de gestión debiendo recurrir a las estructuras de datos para mejorarla.

Hay que recurrir a diferentes mecanismos para simplificar el tamaño de los índices.

- Conversión de texto a minúsculas
- Stemming
- Supresión de palabras vacias
- Compresión de textos

Métodos de RI en el web

El proceso de indización automática puede hacerse:

- Información que proporciona el creador o editor:
  - ✓ Título
  - ✓ Metadatos
- Extrayendo la información directamente del documento.
- Adicionalmente se pueden asignar pesos a los términos según diferentes criterios:
  - ✓ Si aparecen en el título
  - ✓ Según la frecuencia absoluta
  - ✓ Páginas grandes en tamaño
  - ✓ Si aparecen en metadatos

Métodos de RI en el web

Funcionamiento de los motores de búsqueda

- Finalmente, un aspecto importante es el ranking, es decir, el orden en el que se presentan los resultados al usuario, en función de la relevancia de los documentos respecto a la pregunta realizada.

- Esta discriminación por relevancia permite que aparezcan en primer lugar los documentos más relevantes, facilitando el acceso a la información. Se desconoce como se hacen estas tareas en la mayoría de los motores.

Métodos de RI en el web

Funcionamiento de los motores de búsqueda

Existen dos grandes variantes en los algoritmos de ranking:

- ✓ Variantes del modelo vectorial o booleano
- ✓ Los que siguen el principio de extensión de los enlaces

De la primera variante hay tres métodos.

- ★ Booleano extendido
- ★ Vectorial extendido
- ★ Más citado

Métodos de RI en el web

Funcionamiento de los motores de búsqueda

De la segunda variante hay tres métodos.

- WebQuery
- HITS
- PageRank

Métodos de RI en el web

El PageRank

La importancia de una página viene dada por la importancia de las páginas que la enlazan.

$$x_i = \sum_{j \in B_i} \frac{1}{N_j} x_j$$

importancia página i

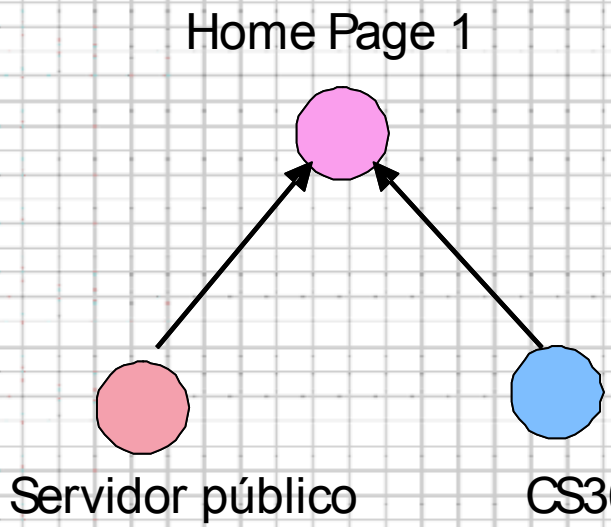
importancia página j

páginas j que enlazan a i

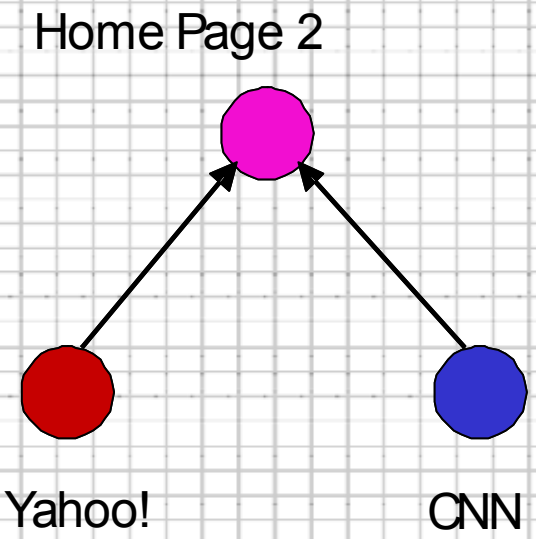
Enlaces salientes de la página j

Métodos de RI en el web

El PageRank



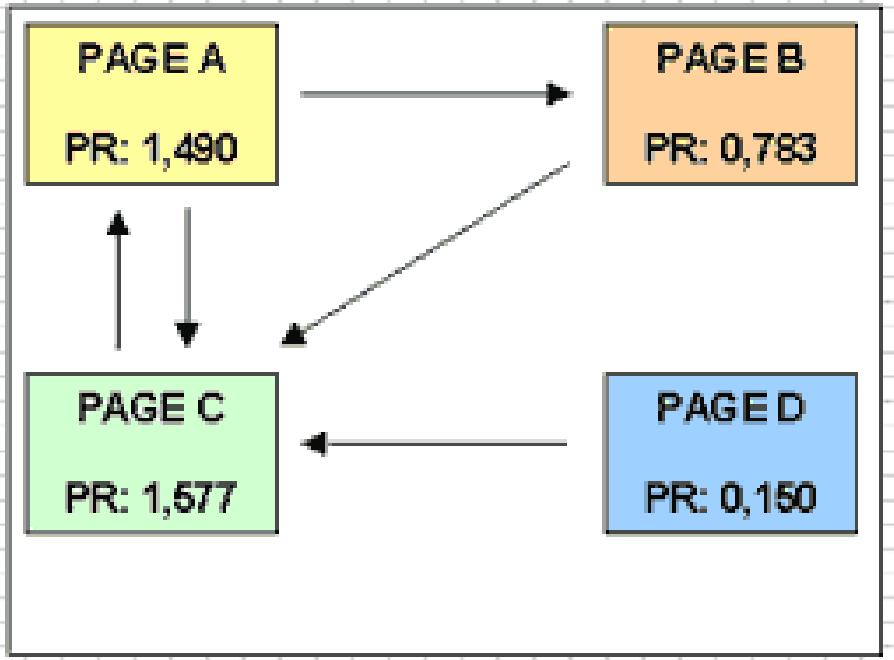
Enlazado por 2
Páginas NO importantes



Enlazado por 2
Páginas importantes

Métodos de RI en el web

El PageRank



reina

Grupo de investigación en **RE**cuperación de la **IN**formación **A**utomatizada. Universidad de Salamanca

INDICE

visite nuestro *web*

http://reina.usal.es