Encontrar documentos a través de las palabras

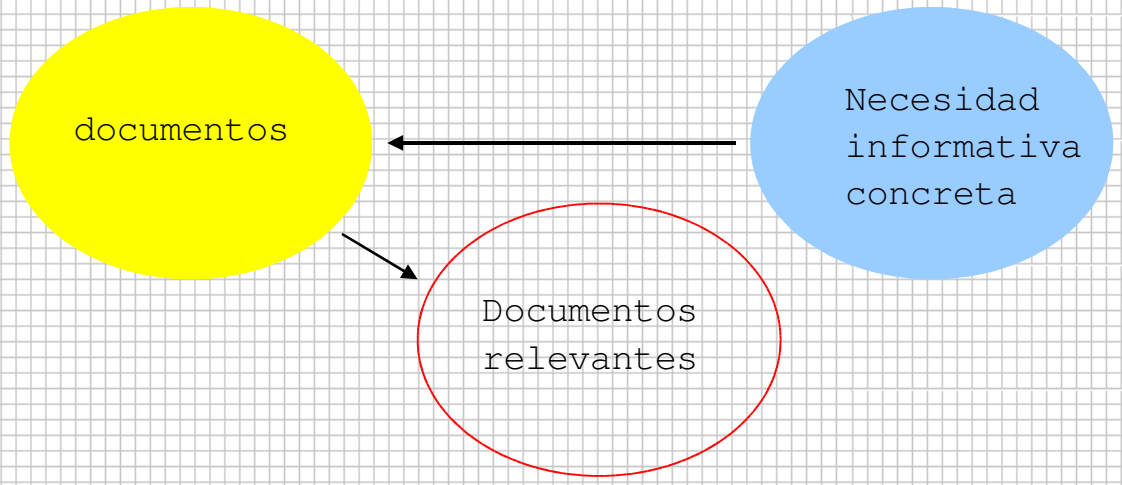
Carlos G. Figuerola
Universidad de Salamanca
Grupo REINA
http://reina.usal.es

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



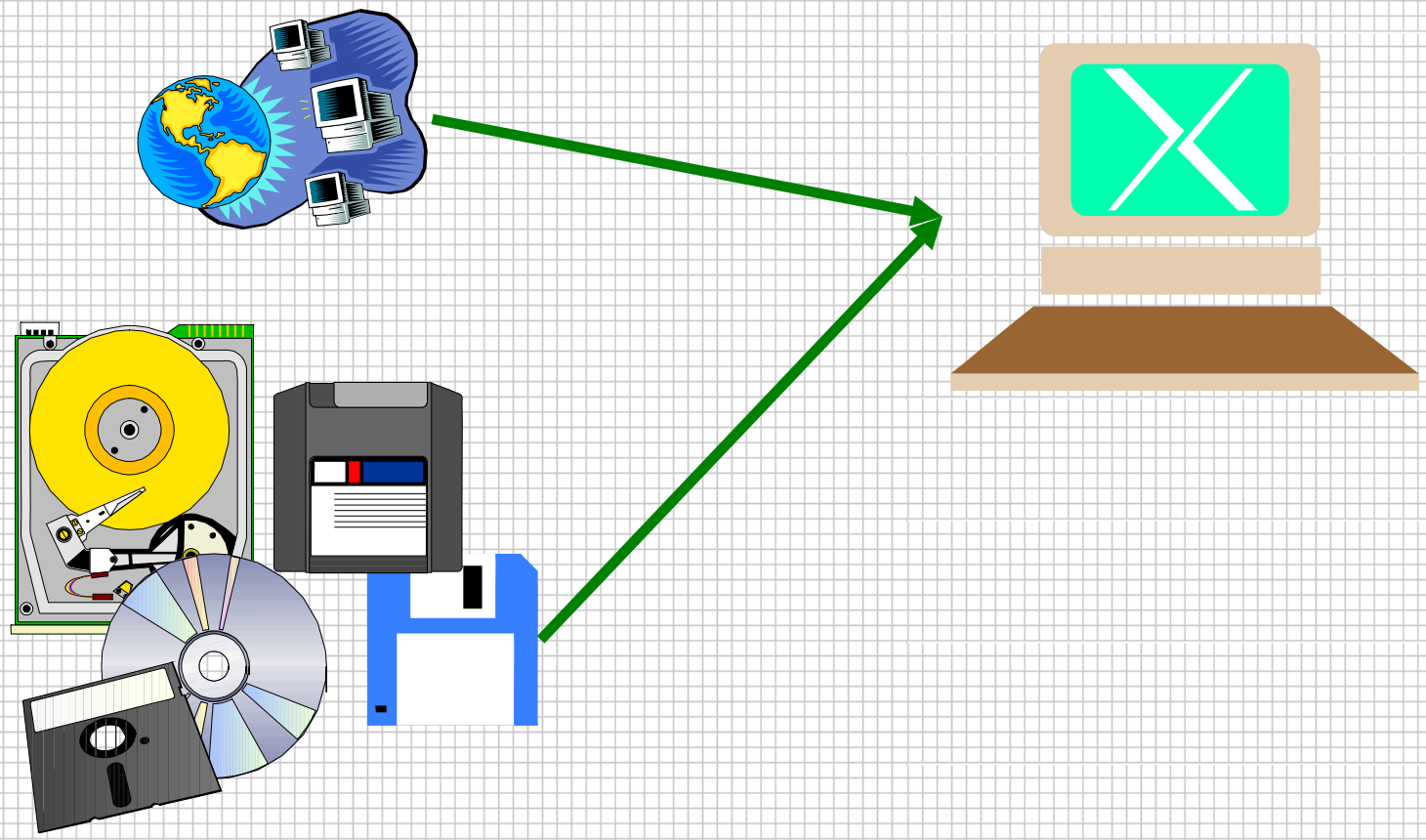
La indización manual

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

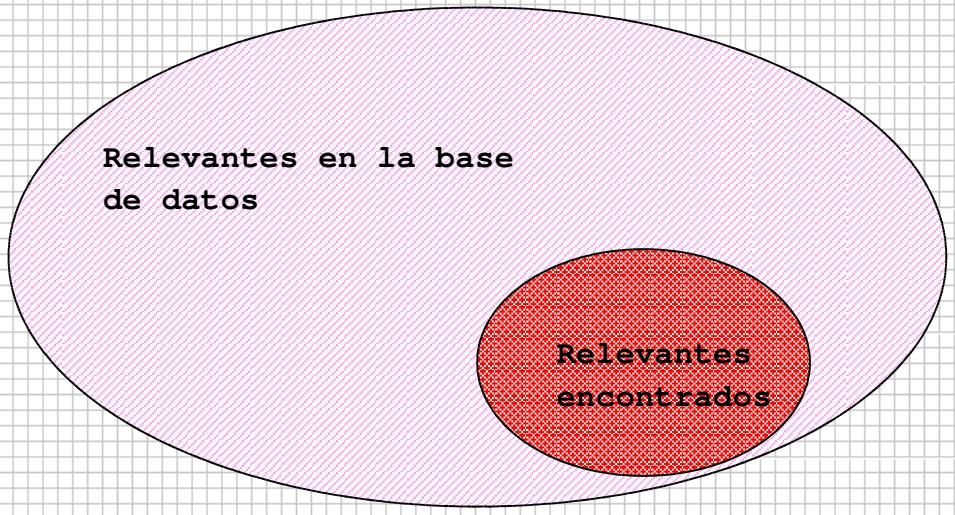
- Después de planteada la necesidad informativa y de obtener los documentos, es necesario evaluar si éstos se corresponden con la necesidad informativa

- 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

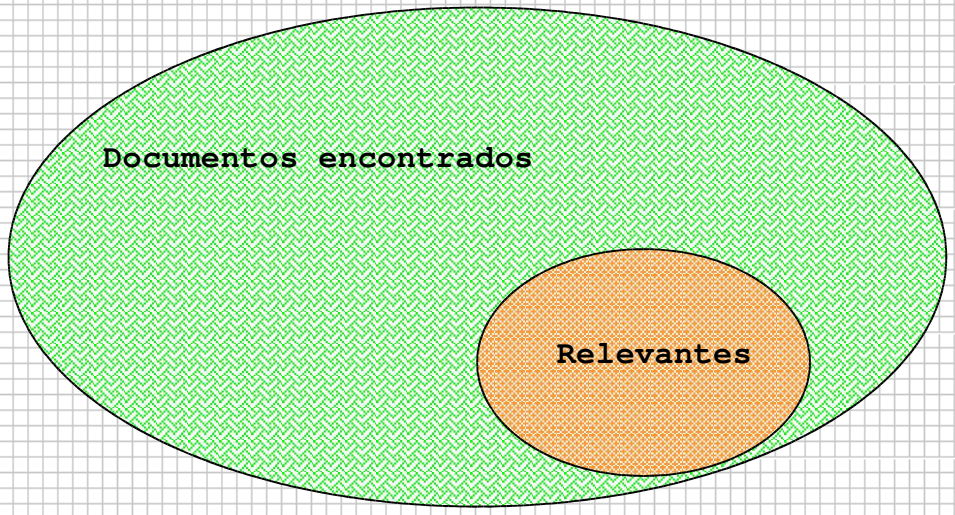
Exhaustividad:
Proporción de documentos relevantes encontrados del total de documentos relevantes en la base de datos



(sólo calculable en bases de datos experimentales)

Evaluación

Precisión:
Proporción de documentos relevantes entre los recuperados

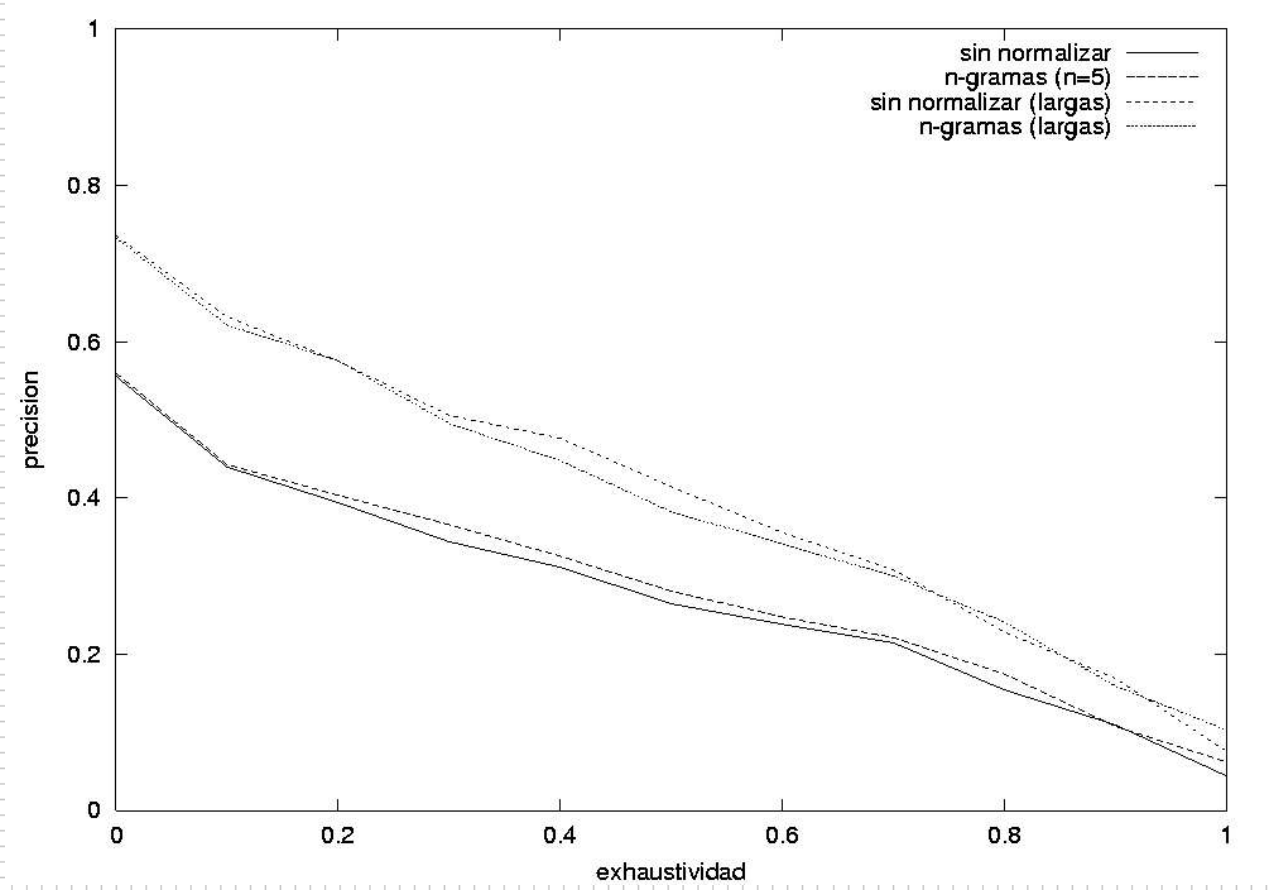


Evaluación. Precisión

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)