Universidad
Carlos III de Madrid

Módulo VI
Modelos de Recuperación

OpenCourseWare

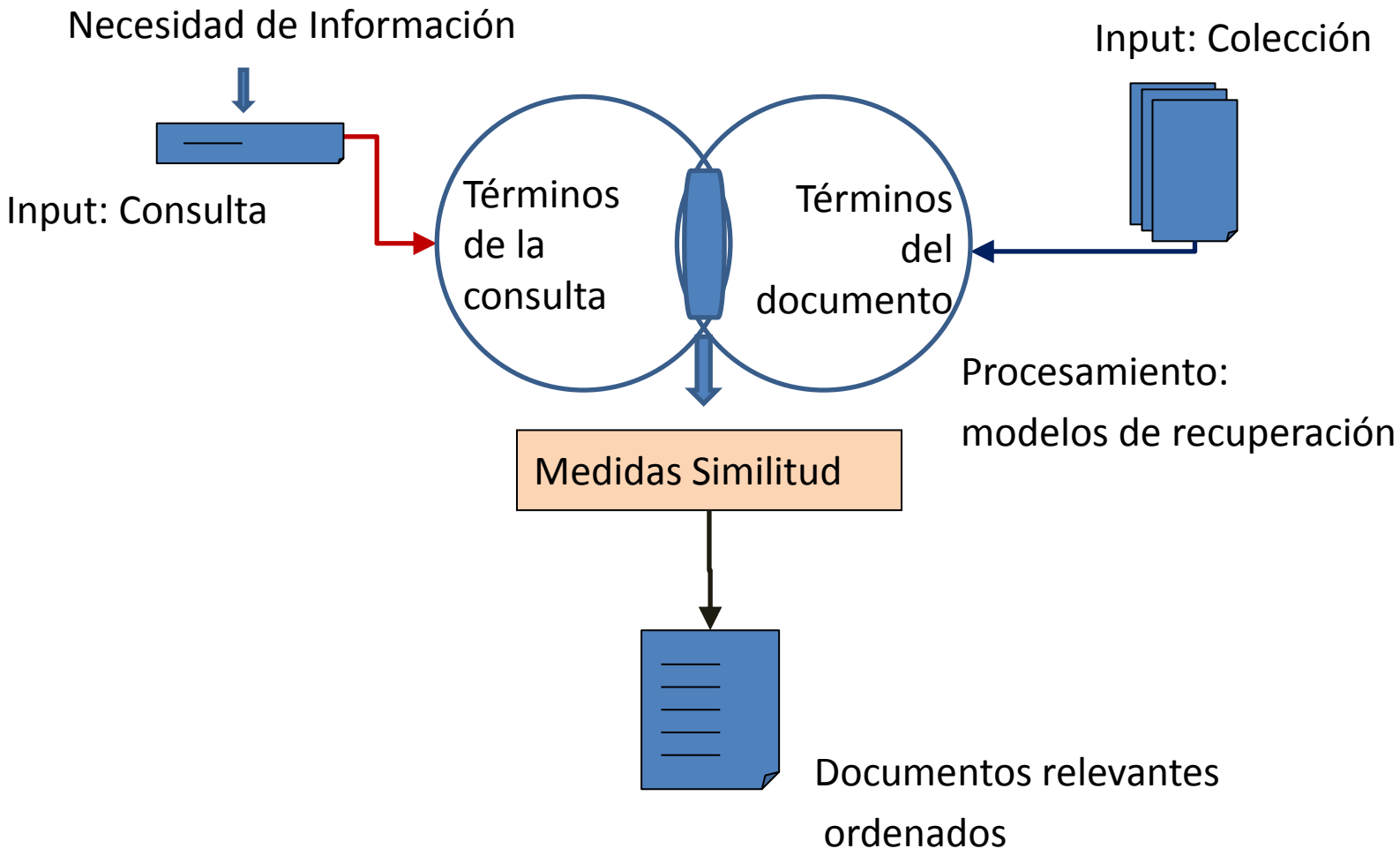
Recuperación y Acceso a la Información

Contenidos

- Conceptos básicos de la recuperación de información

- Modelos clásicos de recuperación
  - Modelo Booleano
  - Modelo Vectorial
  - Modelo Probabilístico

Esquema básico de la evaluación



Conceptos básicos

Dada una colección de documentos.. ¿cómo podemos recuperar la información relevante para nosotros?

- ¿Qué elementos van a ser representativos del contenido de cada documento?
- ¿Cómo vamos a llevar a cabo la representación de ese documento?
- ¿Cómo vamos a llevar a cabo la representación de nuestras consultas?
- ¿Cuál va a ser la función que relaciona una consulta con un conjunto de documentos relevantes?
- ¿Cuál va a ser la relevancia asociada a cada documento?

Representación de documentos y consultas

- Indización: proceso que tiene como objetivo la representación de un documento o consulta mediante los conceptos o palabras que lo componen.
  - Indización manual: conceptos
  - Indización automática: extracción de palabras

- Características:
  - Si utiliza todo el texto se denomina *a texto completo* (full-text)
  - Se eliminan palabras vacías (opcional)
  - Se utilizan sólo sustantivos, verbos y adjetivos (opcional)
  - Se normaliza número y género para obtener taxonomía (opcional)

- En los sistemas de recuperación, la representación documental se realiza con estructuras llamadas **índices inversos**: relación entre términos y documentos que los contienen

Modelos clásicos de Recuperación

- Modelado de Recuperación: función que establece un ranking de documentos, basado en la puntuación que recibe cada documento para una consulta dada

- Modelos clásicos:
  - Booleano
  - Vectorial
  - Probabilístico

- Consideran que los documentos son descritos por una serie de términos K= $\{k_1 \ldots k_n\}$

- Los términos tienen asociados pesos, que varían según el documento que describan. W_{ij}= peso término i en el documento j

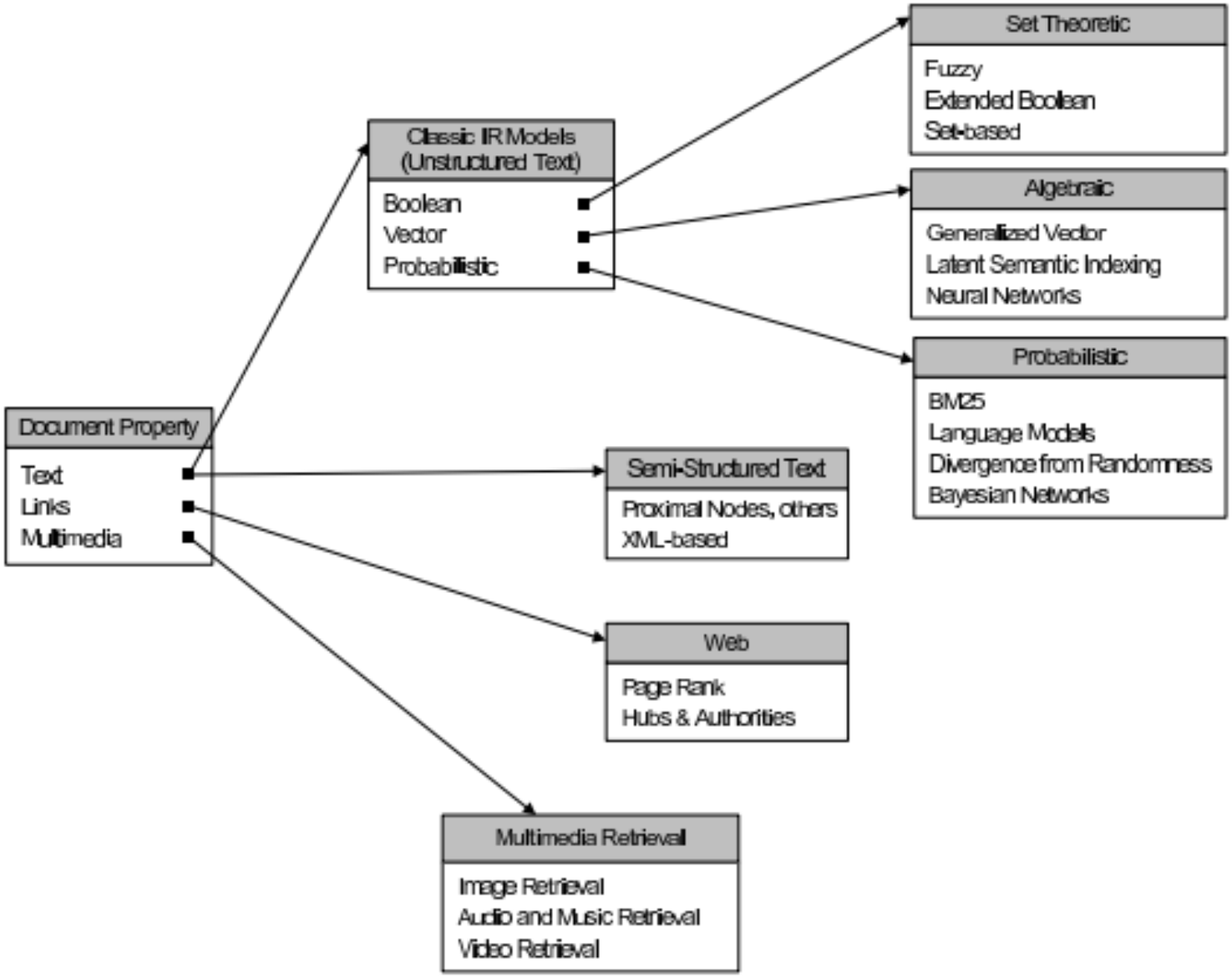
- Un documento es descrito por: $d_j=\{w_{1j} \ldots w_{nj}\}$

Modelo de Recuperación de Información

Es una cuadrupla [D,Q,F,R(q_i,d_j)]

- **D:** representaciones de los documentos de la colección que se desea recuperar
- **Q:** representaciones de las consultas que plasman las necesidades de información del usuario
- **F:** marco que permite establecer una relación entre las representaciones de los documentos y la consulta
- **R(q_i,d_j) o Sim(q_i,d_j):** función de relevancia o similitud que asigna un valor al documento i para una consulta j dada

Clasificación de modelos



Tomado Yates y Neto 2011

Modelo Booleano

- Basado en la lógica de proposiciones (Georges Boole, The Laws of Though, 1854)
- Asigna los pesos en función de si los términos están presentes o no en un documento: pesos binarios
- No hay relevancia parcial: son relevantes o no relevantes
- Las consultas se llevan a cabo en lenguaje booleano:
  - Operadores Booleanos
    - And (+, &, Y, ∧)
    - Or (o, |, ∨)
    - not (no, and not, -, ¬)
    - Además pueden tener paréntesis: precedencia

Modelo Booleano. Ejemplo

- Documento 1: ”los **coches** tienen **ruedas** y circulan por cualquier **vía**”
- Documento 2: ”por la **autopista** pueden circular **coches, motos…**”

Términos (K)	coches	ruedas	Vía	Autopista	motos
Doc. 1	1	1	1	0	0
Doc. 2	1	0	0	1	1

- Q1 : coches AND motos = {D2}
- Q2 : coches OR motos = {D1, D2}
- Q3 : ruedas AND (autopista OR coches) = {D1}

Modelo booleano. Ventajas e inconvenientes

- Ventajas
  - En sistemas de bases de datos relacionales
  - Eficiente y simple
  - Muchos experimentos
  - Álgebra booleana

- Desventajas
  - No ordena por relevancia
  - No tiene en cuenta la frecuencia del término ni su valor discriminante
  - Los usuarios tienen problemas con los operadores booleanos
  - No matching parcial: si cumple algunos AND no lo da como válido
  - Las consultas son simplistas (difícil expresar conceptos como mucho, poco, frecuentemente, …)

Peso de los términos

No todos los términos de un documento tienen el mismo valor discriminante para describir de forma unívoca un documento.

Hay algoritmos para asignar este peso

- Primer Grupo: de, y, la, que, el, en, a, los, no, se, un, las, una, del , pero, lo, por, dijo, con (media=8847, max=22634 min=3126)

- Segundo Grupo: minas, negro, caballo, colinas, duda, estrellas, batalla, piedras, peligro, mal, amigos, aquella, hierba, haya, dio, hojas, viaje, sendero, dama, colina (media=166 max=178 min=156)

- Tercer Grupo: tiembla, tímidos, tiemp, tersa, tiente, tier, tiesamente, tijeretazos, tildado, tapadas, quejoso, quejumbros, quemaban, quemándome, pujanza, pulcros, pulen, pulgadas, pulmón, pululaba (media=1 max=1 min=1)

Peso de los términos: Frecuencia del Término

Luhn: la importancia de un término depende de su frecuencia

Doc.1

A buen amigo, buen abrigo.

Doc.2

A buen vecino, buen amigo

Doc.3

A buen entendedor, pocas palabras

	Doc 1	Doc 2	Doc 3	Total
a	1	1	1	3
buen	2	2	1	4
amigo	1	1		
abrigo	1			
vecino		1		
entendedor			1	
pocas			1	
palabras			1	

Peso de los términos: Frecuencia del Término

- Un modo de establecer el peso de un término (su representatividad) es asignarle su frecuencia de aparición, teniendo en cuenta:
  - tf (term frequency) es la frecuencia de un término i en el documento j.
  - Si un término aparece muchas veces en un documento, su peso en el documento es alto (se puede dividir por max freq o n. terms)

$$tf_{i,j} = f_{i,j}$$

Peso de los términos
Ley de Zipf

F(r)= k / r

k: máxima frecuencia de un término

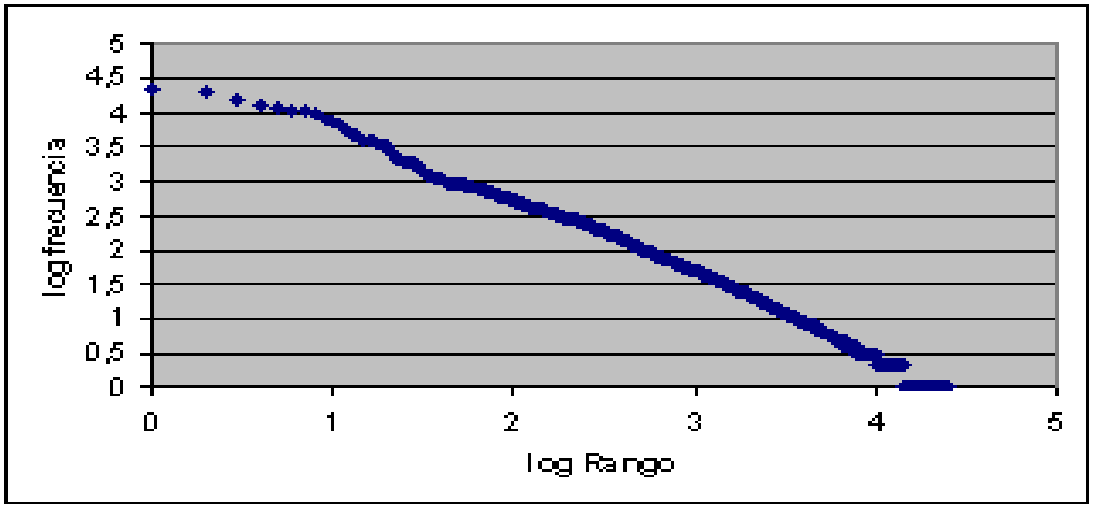
r: rango (orden, de más a menos frecuente)

(ejemplo anterior)

Rango	Palb.	Frec absoluta
1	the	69971
2	of	36411
3	and	28852
4	to	26149
5	a	23237
6	in	21341
7	that	10595
8	is	10049
9	was	9816
10	he	9543

69971/2= =34985 ←

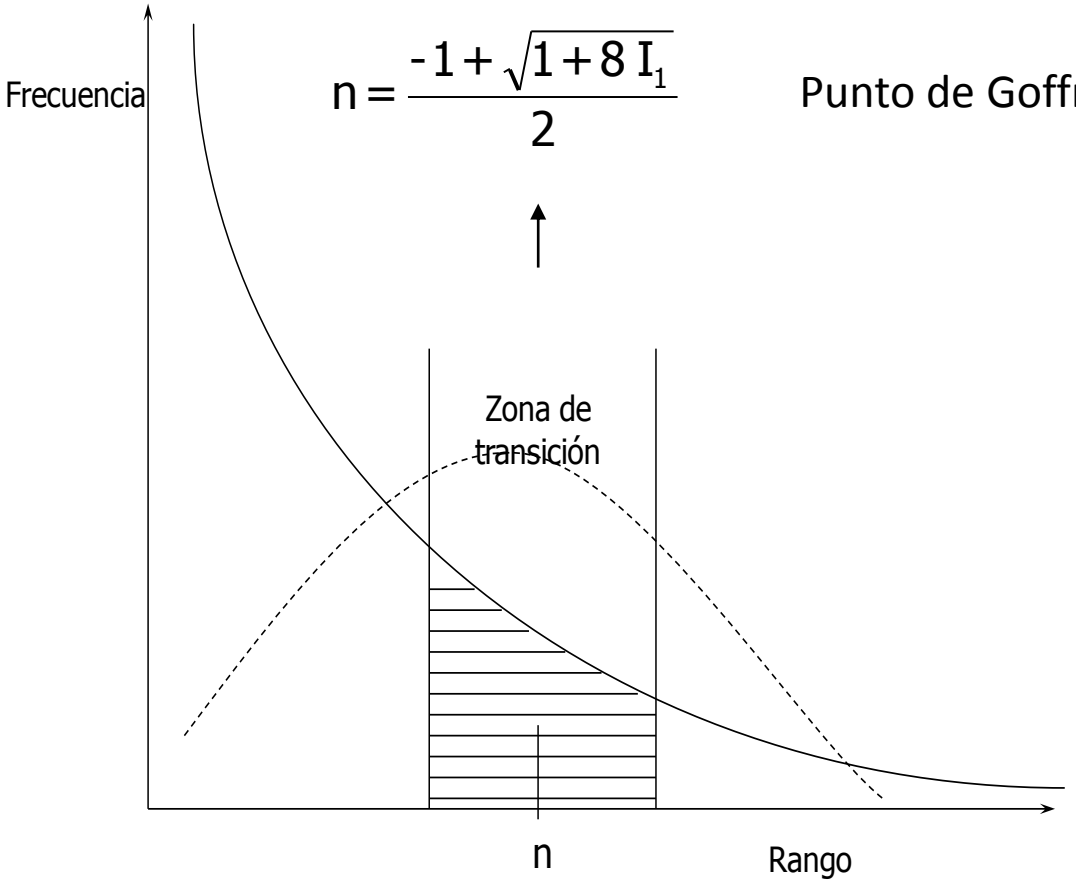
69971/7= =9995 ←



- 25240 términos distintos
- 479212 palabras incluidas repetidas

Peso de términos

I_1: número de palabras en el rango 1 (o número de veces que aparece la palabra más frecuente)



Peso de términos

- Idf (inverse document frequency) es la frecuencia de un término i en el resto de la colección.
- Es una medida de la poder de discriminación de un término para recuperar un documento específico
- Si un término aparece en muchos documentos, su valor discriminativo es pequeño.

$$idf_i = \log (N / n_i)$$

n_i = número de documentos del corpus en los que aparece el término i

N = número total de documentos del corpus

Peso de los términos: Resumen

IDF y TF son los pesos más utilizados para términos

$$idf_i = \log (N / n_i)$$

$$tf_{i,j} = f_{i,j}$$

n_i = número de documentos del corpus en los que aparece el término i

f_{ij} = frecuencia del término i sobre el documento j

N = número total de documentos del corpus

- En la práctica tienden a combinarse ambos. Así el peso del término i en el documento j es igual al producto de ambos:

$$\mathbf{W_{i,j} = tf_{i,j} \times IDF_i = f_{i,j} \times \log (N / n_i)}$$

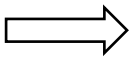
Variaciones de TF-IDF

En las prácticas solo se considerará W_{ij}=tf*IDF=$f_{i,j}$*log (N/n_i)
Aunque lo habitual es el logaritmo en base 2, utilizaremos en base 10

Existen variaciones de TF e IDF (que no usaremos en las prácticas para mayor claridad)

	Peso IDF
Unario	1
Frecuencia inversa	$\log \frac{N}{n_i}$
Frec. Inv. suavizada	$\log(1+\frac{N}{n_i})$
Frec. Inv. Máxima	$\log(1+\frac{max_i n_i}{n_i})$
Frec. Inv. probabilística	$\log \frac{N-n_i}{n_i}$

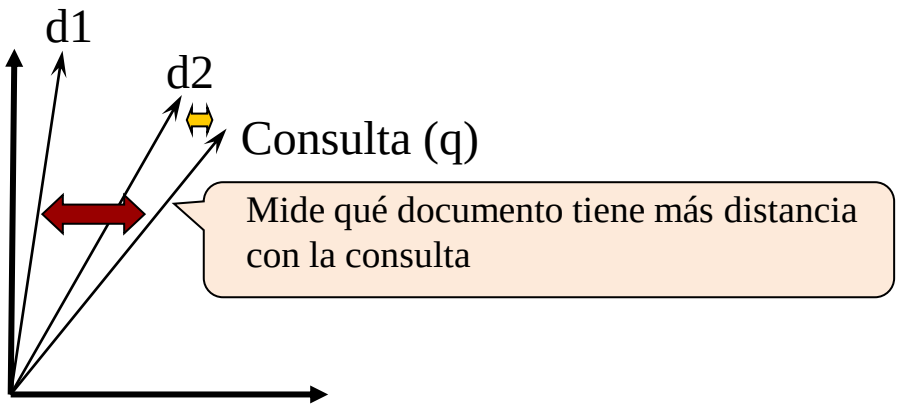
	Peso tf
Binario	{0,1}
Frecuencia	$f_{i,j}$
Normalización logarítmica	$1+\log f_{i,j}$
Normalización 0,5	$0.5+0.5\frac{f_{i,j}}{max_i f_{i,j}}$
Normalizado constante K	$K+(1-K)\frac{f_{i,j}}{max_i f_{i,j}}$



Peso en documento	Peso en la consulta
$f_{i,j} * \log \frac{N}{n_i}$	$(0.5+0.5\frac{f_{i,q}}{max_i \ f_{i,q}}) * \log \frac{N}{n_i}$
$1+\log f_{i,j}$	$\log(1+\frac{N}{n_i})$
$(1+\log f_{i,j}) * \log \frac{N}{n_i}$	$(1+\log f_{i,q}) * \log \frac{N}{n_i}$

Modelo Vectorial

- Cada documento y cada consulta es representada por un vector, con tantas dimensiones como términos en K
- Se asignan pesos positivos y no binarios a los términos
- La similitud entre un documento y una consulta se mide por la distancia existente entre ambos vectores.
- Para el cálculo del coeficiente de similitud se utilizan varias funciones. El más utilizado es el coseno del ángulo entre los dos vectores.



G. Salton et.al., "A Vector Space Model for Automatic Indexing," Communications of the ACM, vol. 18-11, pp. 613–620, 1975.

Modelo Vectorial. Similitud mediante el producto escalar (I)

- $sim(dj,q) = dj \bullet q = \sum_{i=1}^{t} wij \cdot wiq$
  - donde wij es el peso del término i en el documento j y wiq es el peso del término i en la consulta
- Si los pesos son 0 ó 1 el producto escalar es la suma de los términos comunes entre documento y consulta
- Si existe un rango más amplio de pesos, se sumarán los productos de los pesos de los términos en común

Modelo Vectorial. Similitud mediante el producto escalar

- Con pesos binarios (0 ó 1):

En este ejemplo se ha considerado como peso solo tf

	coche	carreter	asiento	mar	multa	motor	rueda
D	1	1	1	0	1	1	0
Q	1	1	0	0	1	0	0

sim(D, Q) =

- Con pesos no binarios:

Supón los siguientes pesos para cada término

$D_1 = (2,3,1,0,2,1,0)$

$D_2 = (3,7,0,0,0,1,1)$

$Q = (1,1,0,0,2,0,0)$

sim(D_1 , Q) =

sim(D_2 , Q) =

Modelo Vectorial. Similitud mediante el producto escalar

- Con pesos binarios (0 ó 1):

En este ejemplo se ha considerado como peso solo tf

	coche	carret	asient	mar	multa	motor	rueda
D	1	1	1	0	1	1	0
Q	1	1	0	0	1	0	0

sim(D, Q) = 3

- Con pesos no binarios:

Supón los siguientes pesos para cada término

$D_1 = (2,3,1,0,2,1,0)$

$D_2 = (3,7,0,0,0,1,1)$

$Q = (1,1,0,0,2,0,0)$

$sim(D_1,Q) = (2\cdot1)+(3\cdot1)+(1\cdot0)+(0\cdot0)+(2\cdot2)+(1\cdot0)+(0\cdot0) = 9$

$sim(D2,Q) = (3\cdot1)+(7\cdot1)+(0\cdot0)+(0\cdot0)+(0\cdot2)+(1\cdot0)+(1\cdot0) = 10$

Modelo Vectorial. Similitud mediante el coseno (I)

En este ejemplo se ha considerado como peso SOLO tf

- Es el producto escalar de ambos vectores normalizado por la longitud de los mismos:

$$CosSim(d_j, q) = \frac{\vec{d_j} \cdot \vec{q}}{\|\vec{d_j}\| \cdot \|\vec{d}\|} = \frac{\sum_{i=1}^{t} w_{ij} \cdot w_{iq}}{\sqrt{\sum_{i=1}^{t} w_{ij}^2} \cdot \sqrt{\sum_{i=1}^{t} w_{iq}^2}}$$

- A diferencia del producto interno, el coseno varía **entre 0 y 1**
- Ejemplo:

D_1 = (2,3,1,0,2,1,0)
D_2 = (3,7,0,0,0,1,1)
Q = (1,1,0,0,2,0,0)

$$\cos Sim(D_1, Q) = 9 / \sqrt{(4+9+1+4+1)\cdot(1+1+4)} = 0.84$$
$$\cos Sim(D_2, Q) = 10 / \sqrt{(9+49+1+1)\cdot(1+1+4)} = 0.53$$

Coseno: D_1 es mejor que D_2
Producto escalar: D_1 es peor que D_2

Nota: Obviamente no se puede calcular para términos con tf=0

Modelo Vectorial. Medidas de similitud

Aunque en las prácticas utilizaremos el coseno, existen alternativas para ver la similitud, que dependiendo del caso pueden ser mejores para el análisis de la semejanza

Similitud		
Producto escalar	$\lvert X \cap Y \rvert$	$\sum_{i=1}^{m} x_i \cdot y_i$
Coeficiente de Dice	$2 \cdot \frac{\lvert X \cap Y \rvert}{\lvert X \rvert + \lvert Y \rvert}$	$\frac{2 \cdot \sum_{i=1}^{m} x_i \cdot y_i}{\sum_{i=1}^{m} x_i^2 + \sum_{i=1}^{m} y_i^2}$
Coseno	$\frac{\lvert X \cap Y \rvert}{\lvert X \rvert \cdot \lvert Y \rvert}$	$\frac{\sum_{i=1}^{m} x_i \cdot y_i}{\sqrt{\sum_{i=1}^{m} x_i^2 \cdot \sum_{i=1}^{m} y_i^2}}$
Coeficiente de Jaccard	$\frac{\lvert X \cap Y \rvert}{\lvert X \rvert + \lvert Y \rvert - \lvert X \cap Y \rvert}$	$\frac{\sum_{i=1}^{m} x_i \cdot y_i}{\sum_{i=1}^{m} x_i^2 + \sum_{i=1}^{m} y_i^2 - \sum_{i=1}^{m} x_i y_i}$

Medidas de similitud

- ¿Qué medida de similitud es más relevante?

Documentos /Términos	A	B
d_1	1.0	1.0
d_2	0.5	1.0
d_3	1.0	0.8
d_4	0.7	0.7
d_5	1.0	0.0
q	1.0	1.0

	Coseno De α	Producto Escalar	Coef. de Dice	Coef. de Jaccard

Tomado de Introducción a la Recuperación de Inf. Tolosa & Bordignon

Medidas de similitud

- Normalmente la mejor Coseno

Documentos /Términos	A	B
d_1	1.0	1.0
d_2	0.5	1.0
d_3	1.0	0.8
d_4	0.7	0.7
d_5	1.0	0.0
q	1.0	1.0

	Coseno De α		Producto Escalar		Coef. de Dice		Coef. de Jaccard	
$d(d_1,q)$	1.00	1º	2.00	1º	2.00	3º	1.00	4º
$d(d_2,q)$	0.95	4º	1.50	3º	2.40	2º	1.75	2º
$d(d_3,q)$	0.99	3º	1.80	2º	2.00	4º	1.84	1º
$d(d_4,q)$	1.00	2º	1.40	4º	2.86	1º	1.58	3º
$d(d_5,q)$	0.71	5º	1.00	5º	2.00	5º	0.50	5º

Tomado de Introducción a la Recuperación de Inf. Tolosa & Bordignon

Peso de los términos (II)
Producto Escalar

- ¿Qué documento es más relevante?

	Termino A	Término B
Documento 1	2	5
Documento 2	8	1
En documentos de la colección (700)	74	12
Consulta (8 términos)	1	1

  - Ocurrencia términos en 1:
  - Ocurrencia términos en 2:
  - Frecuencia inversa término A:
  - Frecuencia inversa término B:
  - TFXIDF D1 =
  - TFXIDF D2 =

Peso de los términos (II) producto escalar con peso TFxIDF

- ¿Qué documento es más relevante?

	Termino A	Término B
Documento 1	2	5
Documento 2	8	1
En documentos de la colección (700)	74	12
Consulta (8 términos)	1	1

- Ocurrencia términos en 1: 2+5 = 7
- Ocurrencia términos en 2: 8+1 = 9->D2 tiene más ocurrencias de los términos de la consulta
- Frecuencia inversa término A: log(700/74)=0,97
- Frecuencia inversa término B: log (700/12)=1,76 -> B es más discriminante que A
- TFXIDF D1 = (1xlog (700/74))(2 xlog (700/74)) + (1xlog(700/12))(5 xlog (700/12)) = **17,49**
- TFXIDF D2 = (1xlog (700/74)) (8 x log (700/74)) + (1 x log (700 /12))(1 x log (700 /12)) = **10,73**

➢ **D1 es más relevante para la consulta según producto escalar y tf-IDF**

Peso de los términos (II)
Producto Vectorial con peso TFxIDF

- ¿Qué documento es más relevante?

	Termino A	Término B
Documento 1	2	5
Documento 2	8	1
En documentos de la colección (700)	74	12
Consulta (8 términos)	1	1

$\text{sim}(Q, D_1) = 0.96$ de forma desarrollada:

$$\frac{(1x\log(700/74))(2x\text{IDF}_A) + (1x\text{IDF}_B)(5\,x\text{IDF}_B)}{\sqrt{(2x\text{IDF}_A)^2 + (5x\text{IDF}_B)^2} \times \sqrt{(1x\text{IDF}_A)^2 + (1x\text{IDF}_B)^2}}$$

$$\text{donde } \text{IDF}_A = \log(700/74);\ \text{IDF}_B = \log(700/12)$$

$\text{sim}(Q, D_2) = 0.67$ de forma desarrollada:

$$\frac{(1x\log(700/74))(8x\text{IDF}_A) + (1x\text{IDF}_B)(1\,x\text{IDF}_B)}{\sqrt{(1x\log(700/74))^2 + (8x\text{IDF}_B)^2} \times \sqrt{(1x\text{IDF}_A)^2 + (1x\text{IDF}_B)^2}}$$

Peso de los términos (II)
Producto Vectorial con peso TFxIDF

- ¿Qué documento es más relevante?

	Termino A	Término B
Documento 1	2	5
Documento 2	8	1
En documentos de la colección (700)	74	12
Consulta (8 términos)	1	1

  - **¿Cómo difieren escalar y vectorial?**

El orden se mantiene frente al escalar.
¿Qué ocurre si hay términos en Documento 1 que no están en la consulta?, ¿varía la fórmula?

Peso de los términos (II)
Producto Vectorial con peso TFxIDF

- ¿Y si hay términos que no están en el documento y no en la consulta o viceversa?

	Término A	Término B	Término C
Documento 1	2	5	4
Documento 2	8	1	0
En documentos de la colección (700)	74	12	20
Consulta (8 términos)	1	1	0

$$\text{sim}(Q, D_1) = 0.78$$

$$\frac{(1\text{xIDF}_A)(2\text{xIDF}_A) + (1\text{xIDF}_B)(5\,\text{xIDF}_B)}{\sqrt{(2\text{xlog}(700/74))^2 + (5\text{xIDF}_B)^2 + (4xIDF_C)^2} \times \sqrt{(1\text{xIDF}_A)^2 + (1\text{xIDF}_B)^2}}$$

$$donde: IDF_C = \log(700/4)$$

El coseno lo penaliza (D1 pasa de 0,96 a 0,78), el escalar no cambia

Modelo vectorial. Ventajas e inconvenientes

- Ventajas
  - Mejores resultados en experimentos, sobre todo con grandes colecciones
  - Tiene en cuenta longitud del documento
  - Grado de relevancia y matching parcial
- Desventajas
  - Tiempo de cálculo (en especial del coseno)
  - El supuesto estadístico de independencia de términos no se cumple
- Extensiones
  - Vectorial generalizado
  - Latent semantic indexing
  - Redes neuronales

Modelo Probabilístico

- Calcula la probabilidad de que un documento sea relevante a una consulta
- Es complejo y costoso computacionalmente
- Necesita un corpus de entrenamiento
- Asume la independencia de términos
  - P(york)=P(york|new) P(pie)=P(pie|apple)
- Asigna pesos a los términos
  - Positivo: es probable que el documento sea relevante
  - Negativo: es probable que el documento no sea relevante
- Muchas alternativas y derivados, como Okapi BM25 y BM25F.

S.E.Robertson y K.S.Jones, "Relevance weighting of search terms", Journal of the American Society for Information Science, vol. 27-3, pp. 129–146, 1976.

Modelo Probabilístico. Cálculo de pesos

$$sim(\vec{d}_j, q) \approx \sum_{i=1}^{t} \left(\log \frac{P(k_i \mid R)}{1 - P(k_i \mid R)} + \log \frac{1 - P(k_i \mid \overline{R})}{P(k_i \mid \overline{R})} \right)$$

$$P(k_i \mid R) = \frac{r_i}{R} \qquad P(k_i \mid \overline{R}) = \frac{n_i - r_i}{N - R}$$

$$\sum_{k_i [q,d_j]} \log \left(\frac{r_i}{R - r_i} \times \frac{N - n_i - R + r_i}{n_i - r_i} \right)$$

Se añade 0,5 por si r_i es muy pequeño

Ecuación (1)
Robertson & Spark-Jones

$$\sum_{k_i [q,d_j]} \log \left(\frac{r_i + 0.5}{R - r_i + 0.5} \times \frac{N - n_i - R + r_i + 0.5}{n_i - r_i + 0.5} \right)$$

N=número de documentos
R=número de documentos relevantes
para la consulta
n=número de documentos con el término
r=número de documentos relevantes con el término

	relevante	No relevante	total
Docs con k	r_i	n_i-r_i	n_i
Docs sin k	R-r_i	N-n_i-(R-r_i)	N-n_i
Todos	R	N-R	N

Tomado de Baeza y Neto, 2011, p. 81-86

Modelo probabilístico

- Sin saber r_i y R no se puede calcular la fórmula. Si los consideramos inicialmente 0

$$sim(d_j, q) \sim \sum_{k_i[q,d_j]} \log\left(\frac{N - n_i + 0.5}{n_i + 0.5}\right)$$

Pero si $n_i > N/2$ saldría negativo, por lo que se usa

$$sim(d_j, q) \sim \sum_{k_i[q,d_j]} \log\left(\frac{N + 0.5}{n_i + 0.5}\right)$$

$$\Pr(k_i \mid R) = 0.5$$

$$\Pr(\overline{k_i} \mid R) = \frac{n_i}{N}$$

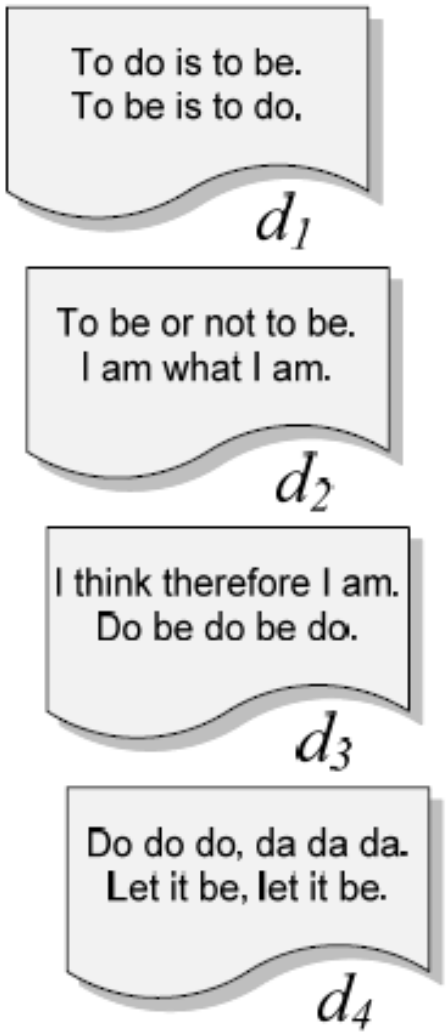
En ausencia de evaluar relevantes se usa esta formula, que es equivalente a IDF

- Cuando se recuperen V documentos iniciales se puede ver cuales son relevantes (umbral o evaluación a mano) así se podrá mejorar el sistema (sustituyendo en Ecuación (1))
  - nota: los docs iniciales pueden estar sesgados
  - V_i es el documento recuperado relevante que contiene el término i.

$$\Pr(k_i \mid R) = \frac{V_i}{V} \qquad \Pr(\overline{k_i} \mid R) = \frac{n_i - V_i}{N - V}$$

Tomado de Baeza y Neto, 2011, p. 81-86

Ejemplo



$$sim(d_j, q) \sim \sum_{k_i[q,d_j]} \log\left(\frac{N+0.5}{n_i+0.5}\right)$$

Ranking para la consulta “to do”

doc	rank computation	rank
d_1	$\log\frac{4+0.5}{2+0.5} + \log\frac{4+0.5}{3+0.5}$	1.210
d_2	$\log\frac{4+0.5}{2+0.5}$	0.847
d_3	$\log\frac{4+0.5}{3+0.5}$	0.362
d_4	$\log\frac{4+0.5}{3+0.5}$	0.362

N=número de documentos=4
n(do)=nº de docs con do=3
n(to)=nº de docs con to=2

Probabilístico con y sin feedback

$$sim(\vec{d}_j, q) \approx \sum_{i=1}^{t} \left(\log \frac{P(k_i \mid R)}{1 - P(k_i \mid R)} + \log \frac{1 - P(k_i \mid \overline{R})}{P(k_i \mid \overline{R})} \right)$$

Inicialmente

$$\Pr(k_i \mid R) = 0.5$$

$$\Pr(\overline{k_i} \mid R) = \frac{n_i}{N}$$

Con feedback (0.5 y 1 son factores de corrección)

$$\Pr(k_i \mid R) = \frac{V_i}{V} + 0.5$$

$$\Pr(\overline{k_i} \mid R) = \frac{n_i - V_i + 0.5}{N - V + 1}$$

Donde: Vi es el número de documentos examinados relevantes que tienen el término i, y V el total de relevantes examinados

	Sin feedback (similar a IDF)	V	Vi	Con feedback
To	$\sum_{i=1}^{t}\left(\log\frac{0.5}{1-0.5}+\log\frac{1-(2/4)}{(2/4)}\right)$	3	2	$\sum_{i=1}^{t}\left(\log\frac{(2/3)}{1-(2/3)}+\log\frac{1-((2-2+0.5)/(4-3+1))}{((2-2+0.5)/(4-3+1))}\right)$
do	$\sum_{i=1}^{t}\left(\log\frac{0.5}{1-0.5}+\log\frac{1-(3/4)}{(3/4)}\right)$	3	1	$\sum_{i=1}^{t}\left(\log\frac{(1/3)}{1-(1/3)}+\log\frac{1-((3-1+0.5)/(4-3+1))}{((3-1+0.5)/(4-3+1))}\right)$

La suma de cada columna sería el resultado para la relevancia del documento para la consulta sin feedback y con feedback

Ejemplo

Q: “oro plata camión”

D1: “envío de **oro** dañado en incendio”

D2: “entrega de **plata** en un **camión** de **plata**”

D3: “envío de **oro** en un **camión**”

D2 y D3 son considerados relevantes

	oro	**plata**	**camión**
N	3	3	3
n	2	1	2
R	2	2	2
r	1	1	2

Probad con los 10 primeros resultados en Internet para ver como se comporta

Modelo Probabilístico. Ventajas en inconvenientes

- Ventajas
  - Ordena los resultados por relevancia
  - Sigue un razonamiento matemático basado en probabilidades, lo que permite que tenga extensiones populares
- Desventajas
  - Poco intuitivo y resultados peores que el vectorial (según Saltón, otros experimentos dicen lo contrario)
  - No es posible conocer al principio el conjunto de documentos relevantes
  - Igual que el vectorial, presupone independencia de términos
  - No normaliza por la longitud del término, no tiene en cuenta frecuencia del término (BM25 que es una variación si)
- Extensiones
  - Considerar la frecuencia de términos en los documentos
  - Redes bayesianas
  - Redes de inferencia bayesianas

Algunos problemas..

- Calcular la similitud mediante la fórmula del coseno de los siguientes documentos (sin tener en cuenta IDF)
  - Doc1: “Demasiado al Este es el Oeste”
  - Doc2: ”Este Sol sale por el Este y se oculta por el Oeste”

- Si se calcula la semejanza entre
  - Doc1: “Un perro ataca a un niño con virus”
  - Doc 2: “Virus ataca al perro de un niño”

- Si se calcula la semejanza entre
  - Doc1: “Tomé vitamina-C”
  - Doc 2: “He tomado Vitamina C”

Referencias alternativas

- Vector Model Information Retrieval http://www.hray.com/5264/math.htm
- Modern Information Retrieval http://www.gib.fi.upm.es/sites/default/files/irmodeling.pdf
- IR Models: The Vector Space Model http://www.csee.umbc.edu/~ian/irF02/lectures/07Models-VSM.pdf
- Modern Information Retrieval. B.Yates, R.Neto, 2011
- Probabilistic model http://nlp.stanford.edu/IR-book/pdf/11prob.pdf
- Introduction to Probabilistic Models for Information Retrieval http://homepages.inf.ed.ac.uk/vlavrenk/doc/pmir-1x2.pdf

Universidad
Carlos III de Madrid

Módulo VI
Modelos de Recuperación

Colaboradores

J.Morato, V.Palacios

M.Marrero, S.Sánchez-Cuadrado, J.Urbano