Gestión de Información en la Web
Máster en Ingeniería Informática
Parte II: Recuperación de información y Sistemas de recomendación

Tema 4: Recuperación de información
4.1. Modelos de RI

Juan Manuel Fernández Luna
Dpto. Ciencias de la Computación e Inteligencia Artificial
jmfluna@decsai.ugr.es

Índice

- Introducción a los Modelos de Recuperación de Información
- Modelos Clásicos
  - Booleano
  - Vectorial
  - Probabilístico

Índice

- Introducción a los Modelos de Recuperación de Información
- Modelos Clásicos
  - Booleano
  - Vectorial
  - Probabilístico

Introducción

- **Recuperación de Información (RI)** :

  Disciplina encargada de la representación, almacenamiento y organización de la información, y su posterior acceso y recuperación para responder a las necesidades de un usuario.

- Un **Modelo de RI** es la especificación sobre cómo representar documentos y consultas, y cómo comparar unos y otras

- El objetivo es obtener un orden (*ranking)* de los documentos recuperados que refleje la relevancia de los documentos a la consulta del usuario

Introducción

- SIMILARIDAD es el concepto básico en R.I.

  los documentos que utilizan vocabulario similar tienden a ser relevantes a las mismas consultas.

- Se pueden utilizar distintos criterios para medir la similaridad entre documentos y consultas
  - igual conjunto de términos,
  - comparten términos con peso,
  - versosimilitud de relevancia

- Cada criterio nos lleva a un modelo de recuperación de información distinto

Bolsa de Palabras (Bag of words)

- Similaridad no tiene en cuenta el orden: es una forma efectiva de abordar la problemática de la R.I.
  - Comparan palabras independientemente del orden en que aparecen en el texto:

- Por ejemplo, consideremos los siguientes ordenes:
  - Aleatorio:

  palabras orden aparecen texto comparan independientemente
  - Alfabético

  aparecen comparan independientemente orden palabras texto
  - Real

  Comparan palabras independientemente orden aparecen texto

Ejemplo (obtenido de James Allan)

¿De qué trata este documento?

- 16 × said
- 14 × McDonalds
- 12 × fat
- 11 × fries
- 8 × new
- 6 × company french nutrition
- 5 × food oil percent reduce taste Tuesday
- 4 × amount change health Henstenburg make obesity
- 3 × acids consumer fatty polyunsaturated US
- 2 × amounts artery Beemer cholesterol clogging director down eat estimates expert fast formula impact initiative moderate plans restaurant saturated trans win
- 1 × added addition adults advocate affect afternoon age Americans Asia battling beef bet brand Britt Brook Browns calorie center chain chemically … crispy customers cut … vegetable weapon weeks Wendys Wootan worldwide years York

Ejemplo (obtenido de James Allan)

El texto original dice

- **McDonald's slims down spuds**

  Fast-food chain to reduce certain types of fat in its french fries with new cooking oil.
- **NEW YORK (CNN/Money) - McDonald's Corp. is cutting the amount of "bad" fat in its french fries nearly in half, the fast-food chain said Tuesday as it moves to make all its fried menu items healthier.**

  But does that mean the popular shoestring fries won't taste the same? The company says no. "It's a win-win for our customers because they are getting the same great french-fry taste along with an even healthier nutrition profile," said Mike Roberts, president of McDonald's USA.

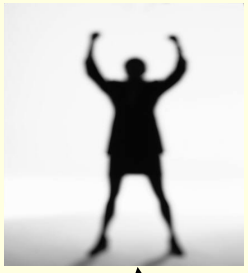
  But others are not so sure. McDonald's will not specifically discuss the kind of oil it plans to use, but at least one nutrition expert says playing with the formula could mean a different taste.

  Shares of Oak Brook, Ill.-based McDonald's (MCD: down $0.54 to $23.22, Research, Estimates) were lower Tuesday afternoon. It was unclear Tuesday whether competitors Burger King and Wendy's International (WEN: down $0.80 to $34.91, Research, Estimates) would follow suit. Neither company could immediately be reached for comment.

  ...

http://money.cnn.com/2002/09/03/news/companies/mcdonalds/index.htm

Arquitectura de SRI simple



Necesidad Información

Representación

Consulta

Evaluación/ Realimentación

Colección Documental

Representación

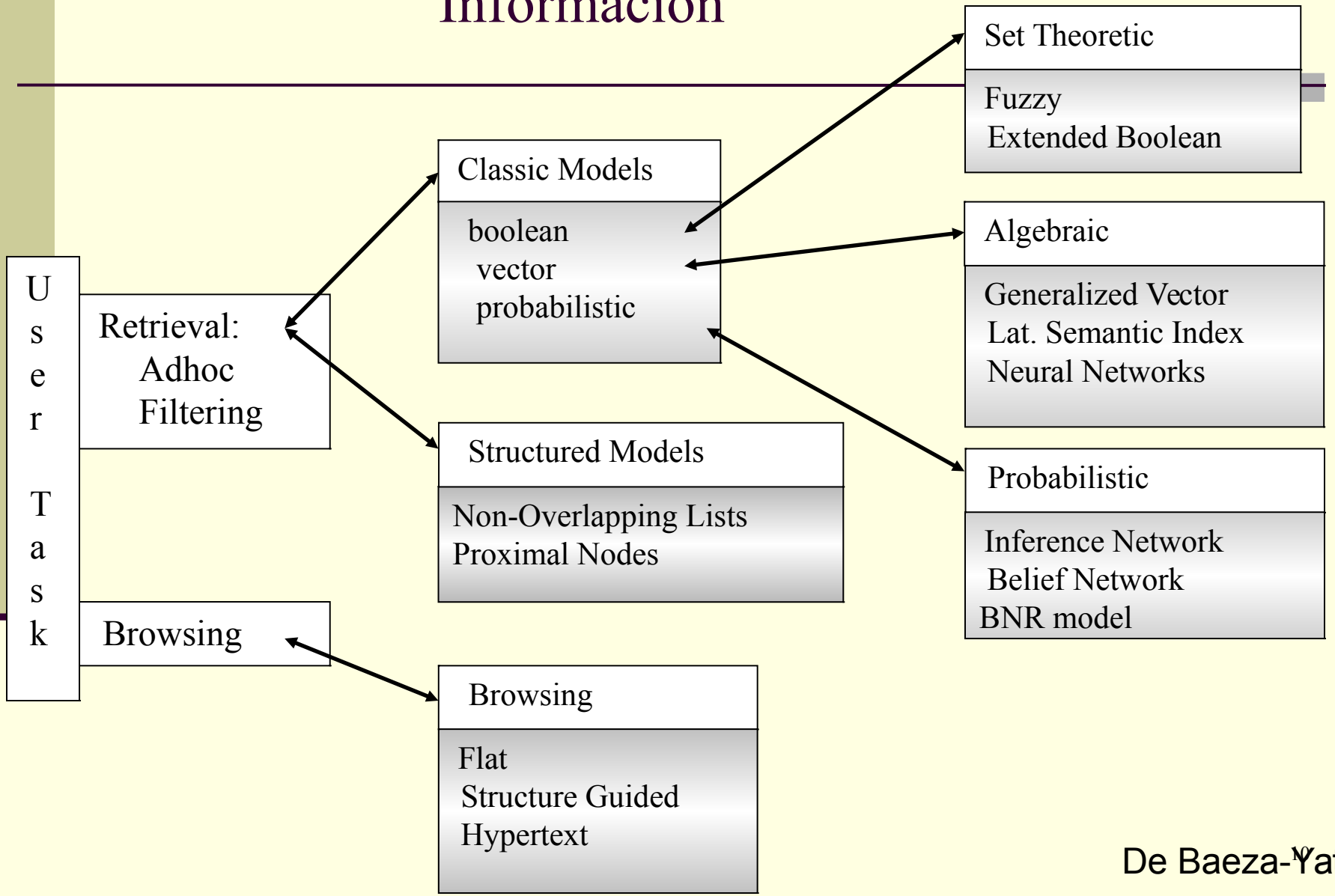
Fichero Invertido

Comparación

Documentos recuperados

Ordenados por relevancia

Modelos de Recuperación de Información



Índice

- Introducción a los Modelos de Recuperación de Información
- Modelos Clásicos
  - Booleano
  - Vectorial
  - Probabilístico

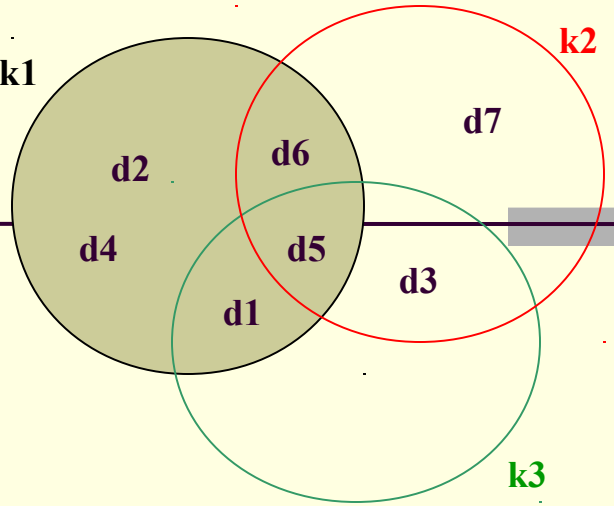
Modelo Booleano

- Es el modelo más simple, basado en teoría de conjuntos
- Documentos se representan como conjunto de términos
  - cada término toma dos valores
    (presente 1/ausente 0) el el documento
- Consultas se representan como expresiones booleanas
  - Conjunto de términos relacionados con conectores
    AND, OR, NOT.
  - formalismo claro y semántica precisa
  - $q = ka \wedge (kb \vee \neg kc)$

  Se transforma en la forma normal disyuntiva

  $(ka \wedge kb \wedge kc) \vee (ka \wedge kb \wedge \neg kc) \vee (ka \wedge \neg kb \wedge \neg kc)$
  - La consulta original se transforma en consultas menores que se pueden lanzar de forma independiente y el resultado final se obtiene mezclado las distintas salidas.

El modelo Booleano: Ejemplo



	k1	k2	k3	consulta
d1	1	0	1	
d2	1	0	0	Q1
d3	0	1	1	
d4	1	0	0	Q1
d5	1	1	1	Q0,Q1
d6	1	1	0	Q1
d7	0	1	0	Q2
Q0 = k1 ∧ k2 ∧ k3	Q1=k1 ∧ (k2 ∨¬k3)		Q2=¬k1 ∧ k2 ∧ ¬k3	

Desventajas del Modelo Booleano

- El mundo no es ni blanco ni negro, también hay grises: Recuperación está basada en criterios binarios, sin dar opción a un emparejamiento parcial
- No presenta los documentos por orden de relevancia (no puede proporcionar un grado de relevancia)
- Los usuarios pueden tener problema a la hora de especificar la consulta: La necesidad de información debe ser expresada por una expresión booleana.

  Por tanto, las consultas tienden a ser muy simples
- Por tanto, o bien un SRI booleano devuelve demasiados o muy pocos documentos al usuario.

Índice

- Introducción a los Modelos de Recuperación de Información
- Modelos Clásicos
  - Booleano
  - Vectorial
  - Probabilístico

Modelo Vectorial

Idea: Utilizar pesos no binarios para indexar tanto los documentos como las consultas

- **Documento d**: representado por una secuencia de términos (vector de pesos)

  $d = (\omega(1), \omega(2), \omega(3), \ldots, \omega(|t|))$

  Con $\omega(t)$ el peso del término t en el documento (0 si t no indexa el documento)

- **Consulta q**: representada por una secuencia de términos

  $q = (\omega q(1), \omega q(2), \omega q(3), \ldots, \omega q(|t|))$

  Con $\omega q(t)$ el peso del término t en el la consulta (0 si t no indexa la consulta)

- El sistema es un espacio vectorial |t|-dimensional, siendo |t| el número de términos en la colección.

Modelo Vectorial: Ejemplo

consulta		
q	*web graph*	
documentos	texto	términos
d_1	*The web … a web graph ….*	*web graph*
d_2	*Is a graph … web … net … graph … net*	*graph web net*
d_3	*A complex web page …*	*page web complex*

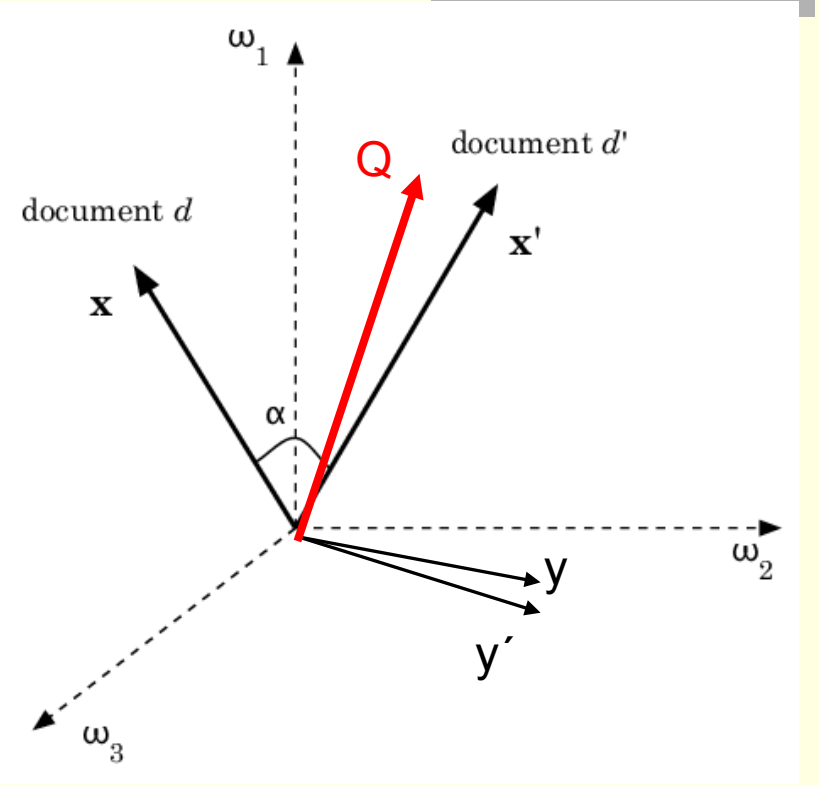
	web	graph	net	page	complex
q	wq1	wq2	0	0	0
D1	W11	W12	0	0	0
D2	W21	W22	W23	0	0
D3	W31	0	0	w34	w35

1.- ¿Cómo comparar consulta y documentos?

2.-¿Qué pesos nos permiten recuperar más documentos?

Modelo Vectorial : Similaridad entre vectores

- Medida de Similaridad: En un espacio vectorial el coseno del angulo entre dos vectores.
- Los documentos con muchos términos en comun tendrán vectores más cercanos (y e y´) que aquellos con pocos términos comunes (x,x´)



w1,w2, w3 son términos

Modelo Vectorial: Medida Coseno

- Similaridad entre vectores:

  El coseno del angulo formado por los vectores *x* y q es

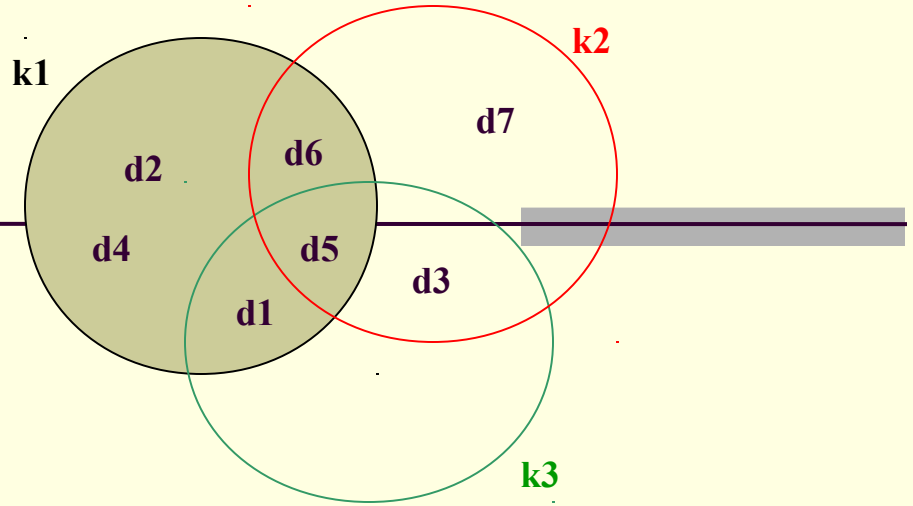
$$\cos(x,q) = \frac{x^T q}{\|x\| \cdot \|q\|}$$

$$\cos(x,q) = \frac{\sum_{i=1}^{t} w_{i,x} \times w_{i,q}}{\sqrt{\sum_{i=1}^{t} w_{i,x}^2} \times \sqrt{\sum_{i=1}^{t} w_{i,q}^2}}$$

Propiedades:

- 0<= cos(x,q) <=1
- Devuelve un documento, aunque satisfaga parcialmente la consulta

Modelo Vectorial
Ejemplo II



	k1	k2	k3	q • dj
d1	1	0	1	4
d2	1	0	0	1
d3	0	1	1	5
d4	1	0	0	1
d5	1	1	1	6
d6	1	1	0	3
d7	0	1	0	2
q	1	2	3	

Modelo Vectorial: Cálculo de pesos

- ¿Cómo calcular los pesos *wix y wiq* ?
- Un buen criterio es tener en cuenta dos efectos:
  - Importancia del término en el documento (tf): Un término que aparece muchas veces en un documento es más importante que otro que sólo aparece una vez
    - *tf frecuencia del término en el documento*
  - Importancia del término en la colección (idf): un término que aparece pocas veces en la colección tiene un mayor poder discriminador que un término que aparezce en todos los documentos.
    - *idf (inverse document frequency) freq. documental inversa*

Modelo Vectorial: Cálculo de pesos

tf Frecuencia del término

$$TF_{ix} = \frac{n_{ix}}{\max_l n_{l,x}}$$

idf *Frecuencia documental inversa*

$$IDF_i = \log \frac{N}{n_i}$$

- *pesos finales:*

$$w_{i,x} = TF_{i,x} \times IDF_i$$

$$w_{i,q} = \left(0.5 + \frac{0.5 \times TF_{i,q}}{\max_l TF_{l,q}} \right) \times IDF_i$$

Modelo Vectorial

- El modelo vectorial con pesos *tf-idf* representa una buena estrategia en colecciones generales.
- El modelo vectorial es competitivo frente a otras alternativas. Es simple y fácil de calcular
- Ventajas:
  - mejora la calidad recuperadora del sistema
  - permite recuperar documentos que sólo se emparejan parcialmente con las consultas.
  - permite ordenar los documentos por grado de similaridad con respecto a la consulta.
- Desventaja:
  - asume independencia de términos

Índice

- Introducción a los Modelos de Recuperación de Información
- Modelos Clásicos
  - Booleano
  - Vectorial
  - Probabilístico

Modelos de R.I. Probabilísticos

- El objetivo es abordar la problemática de la RI utilizando el formalismo probabilístico.
- Asume que dada una consulta del usuario, hay una respuesta *ideal.*
- Una consulta es una especificación de las propiedades de dicha respuesta ideal.
- Los modelos probabilísticos, siendo de los más antiguos, son hoy día unos de los más estudiados.
  - *Los modelos tradicionales se basaban en ideas claras, pero no solían ganar en comportamiento. Actualmente la situación ha cambiado.*

Modelos de R.I. Probabilísticos

- Razones para usar la probabilidad:
  - La disciplina de la RI esta plagada de incertidumbre:
    - En la representación del contenido de los documentos (indexación).
    - En la descripción de la necesidad de información del usuario (consulta).
  - La probabilidad proporciona una base teórica sólida para el manejo de incertidumbre y por tanto para el diseño de sistemas de R.I.

Modelos de R.I. Probabilísticos

Para cada documento y consulta, tratan de responder a la siguiente pregunta:

¿*Cuál es la probabilidad de que este documento sea relevante a esta consulta*?

- Se basa en el *Probability Ranking Principle:*

  "La efectividad global de un sistema es la mejor posible cuando los documentos se presentan en orden creciente de la probabilidad de relevancia (donde las probabilidades se estiman de la forma más precisa posible)"

Modelos de R.I. Probabilísticos

Problema:

¿Cómo se determinan los valores de probabilidad necesarios?

Conceptos básicos:

"Para una consulta dada, si conocemos que algunos documentos son relevantes, los términos que ocurren en esos documentos deben tener un mayor peso a la hora de buscar otros documentos relevantes.

Haciendo suposiciones sobre la distribución de los términos y aplicando el Teorema de Bayes es posible estimar los pesos de forma teórica."

Van Rijsbergen

Modelos de R.I. Probabilísticos

- OKAPI: Es el nombre dado a una familia de SRI experimentales basados en el modelo probabilístico de Robertson-Spark Jones.

- Considera hasta cinco factores distintos:
  - Frecuencia en la colección
  - Frecuencia en el documento
  - Información de relevancia
  - Longitud del documento
  - Frecuencia en la consulta

  Estos factores se combinan para dar un peso a cada par término-documento.

  Finalmente estos valores individuales se combinan para dar un peso final que nos diga cómo de relevante es el documento a la consulta.

Modelos de R.I. Probabilísticos

Sea d un documento de la colección.
Sea *Rel* el suceso "El usuario juzga d como **relevante** para la consulta" y
sea *NRel* el suceso "El usuario lo juzga como **no-relevante.**

Objetivo: Calcular P(Rel|d), pero en su lugar calcularemos la razón con p(NRel|d), pues preserva el mismo orden (PRP):

$$sim(q,d) = \frac{P(\mathrm{Rel}|d)}{P(\overline{\mathrm{Rel}}|d)}$$

Modelos de R.I. Probabilísticos

- Aplicanto el Teorema de Bayes

$$P(\text{Rel}|d) = \frac{P(d|\text{Rel}) \cdot P(\text{Rel})}{P(d)}$$

$$P(\overline{\text{Rel}}|d) = \frac{P(d|\overline{\text{Rel}}) \cdot P(\overline{\text{Rel}})}{P(d)}$$

$$sim(q,d) = \frac{P(\text{Rel}|d)}{P(\overline{\text{Rel}}|d)} = \frac{\frac{P(d|\text{Rel}) \cdot P(\text{Rel})}{P(d)}}{\frac{P(d|\overline{\text{Rel}}) \cdot P(\overline{\text{Rel}})}{P(d)}}$$

- P(d|Rel) = probabilidad de que, sabiendo que hemos escogido un documento relevante, ese documento sea d. P(d|NRel) es la recíproca; es decir, sabiendo que hemos escogido un documento no relevante, la probabilidad de que ese documento sea d.
- P(Rel) y P(NRel) son las probabilidades de escoger un documento de la colección al azar y que sean relevante e irrelevante, respectivamente.
- P(d) es la probabilidad de seleccionar d al azar de entre toda la colección de documentos.

Modelos de R.I. Probabilísticos

Independencia de los atributos: Binary Independence Model.

- Los pesos de los términos que definen a un documento son binarios, es decir, un término únicamente podrá estar presente o ausente en el documento, no contemplándose otros posibles valores (este concepto también es usado en el modelo booleano).
- Los términos que definen un documento son independientes entre sí. Esta suposición ya fue considerada en el modelo vectorial.

$$P(d|\mathrm{Rel}) = \prod_{i=1}^{M} P(w_{i,d}|\mathrm{Rel})$$

$$P(d|\overline{\mathrm{Rel}}) = \prod_{i=1}^{M} P(w_{i,d}|\overline{\mathrm{Rel}})$$

$$sim(q,d) \approx \prod_{i=1}^{M} \frac{P(w_{i,d}|\mathrm{Rel})}{P(w_{i,d}|\overline{\mathrm{Rel}})}$$

Modelos de R.I. Probabilísticos

Notamos p_i como la probabilidad de que el término i esté presente en un documento relevante y q_i como la probabilidad de que el término i esté presente en un documento irrelevante:

$$p_i \equiv P\left(w_{i,d}=1 \middle| \text{Rel}\right),\ 1-p_i \equiv P\left(w_{i,d}=0 \middle| \text{Rel}\right)$$
$$q_i \equiv P\left(w_{i,d}=1 \middle| \overline{\text{Rel}}\right),\ 1-q_i \equiv P\left(w_{i,d}=0 \middle| \overline{\text{Rel}}\right)$$

Y dividir el cálculo de la similitud en dos partes, en función de si el término está presente o ausente en el documento d:

$$sim(q,d) \approx \prod_{w_{i,d}=1} \frac{p_i}{q_i} \cdot \prod_{w_{i,d}=0} \frac{1-p_i}{1-q_i}$$

Para simplificar la notación, vamos a denotar como $x_i = 1$ si el término i (t_i) aparece en el documento ($x_i = 0$ si no aparece) e, $y_i = 1$ si t_i aparece en la consulta q ($y_i = 0$ en caso contrario).

Modelos de R.I. Probabilísticos

	t_i aparece en d	t_i no aparece en d
t_i aparece en q	$x_i = y_i = 1$	$x_i = 0, y_i = 1$
t_i no aparece q	$x_i = 1, y_i = 0$	$x_i = y_i = 0$

$$sim(q,d) \approx \prod_{x_i=1} \frac{p_i}{q_i} \cdot \prod_{x_i=0} \frac{1-p_i}{1-q_i}$$

Modelos de R.I. Probabilísticos

Todos los términos que no están presentes en la consulta son innecesarios para el cálculo de la relevancia.

Ello se traduce matemáticamente diciendo que la probabilidad de que esté presente uno de estos términos en un documento relevante es la misma que la de que esté ausente.

En notación de probabilidades, eso implica que $p_i = q_i$, si t_i no aparece en q (es decir, si $y_i = 0$).

En base a esto, podemos derivar el cálculo de las probabilidades en función de si los términos aparecen o no en el documento y aparecen o no en la consulta:

$$sim(q,d) \approx \prod_{x_i=1} \frac{p_i}{q_i} \cdot \prod_{x_i=0} \frac{1-p_i}{1-q_i} =$$

$$= \prod_{x_i=y_i=1} \frac{p_i}{q_i} \cdot \prod_{x_i=1, y_i=0} \frac{p_i}{q_i} \text{¿} \prod_{x_i=y_i=0} \frac{1-p_i}{1-q_i} \text{¿} \prod_{x_i=0, y_i=1} \frac{1-p_i}{1-q_i}.$$

Modelos de R.I. Probabilísticos

En el caso de que los términos no estén presentes en la consulta (es decir, cuando $y_i = 0$), podemos considerar $p_i = q_i$ y, en consecuencia, eliminar el segundo y tercer término de la fórmula (su valor es uno). Por lo tanto, el cálculo de la similitud quedaría como:

$$sim(q,d) \approx \prod_{x_i = y_i = 1} \frac{p_i}{q_i} \cdot \prod_{x_i = 0, y_i = 1} \frac{1 - p_i}{1 - q_i}$$

Introduciendo $\prod_{x_i = y_i = 1} \frac{1 - p_i}{1 - q_i} \cdot \prod_{x_i = y_i = 1} \frac{1 - q_i}{1 - p_i}$

$$sim(q,d) \approx \prod_{x_i = y_i = 1} \frac{p_i}{q_i} \cdot \prod_{x_i = 0, y_i = 1} \frac{1 - p_i}{1 - q_i} \text{¿} \prod_{x_i = y_i = 1} \frac{1 - p_i}{1 - q_i} \text{¿} \prod_{x_i = y_i = 1} \frac{1 - q_i}{1 - p_i} =$$

$$= \prod_{x_i = y_i = 1} \frac{p_i (1 - q_i)}{q_i (1 - p_i)} \cdot \prod_{y_i = 1} \frac{1 - p_i}{1 - q_i} \cdot$$

Modelos de R.I. Probabilísticos

Si nos centramos únicamente en aquellos términos que nos permiten discriminar a los documentos relevantes para una consulta determinada, la similitud se puede calcular mediante:

$$sim(q,d) \approx \sum_{x_i = y_i = 1} c_i + k$$

donde, c_i es lo que se denomina grado de relevancia o valor del estado de recuperación (RSV – *Retrieval Status Value*):

$$c_i = \log \frac{p_i(1-q_i)}{q_i(1-p_i)}$$

- Es importante destacar que el RSV se calcula en base a los términos que aparecen tanto en la consulta realizada como en el documento analizado.
- Por otra parte, k toma un valor constante para cada consulta, ya que se calcula sobre todos los términos de la consulta, independientemente de si aparecen o no el documento, por lo que se puede eliminar.

Modelos de R.I. Probabilísticos

Así, la similitud entre una consulta y un documento se calcula como:

$$sim(q,d) \approx \sum_{x_i=y_i=1} c_i = \sum_{x_i=y_i=1} \log \frac{p_i(1-q_i)}{q_i(1-p_i)}$$

Inicialmente se necesitan unas estimaciones de valores q_i y p_i. La probabilidad de que t_i esté presente en un documento relevante (p_i) se considera constante para todos los términos de la consulta. La probabilidad de que t_i esté presente en un documento no relevante (q_i) se puede aproximar mediante la distribución del término en los documentos de la colección

$$p_i = 0{,}5,$$

$$q_i = \frac{n_i}{N}$$

$$c_i = \log \frac{0{,}5 \cdot \left(\frac{N-n_i}{N}\right)}{\frac{n_i}{N} \cdot 0{,}5} = \log \frac{N-n_i}{n_i}$$

Muy similar al idf del vectorial → sumar los idf de los términos simultáneos de la consulta y del Documento.

Modelos de R.I. Probabilísticos

Para el resto de iteraciones, asumiendo que se han revisado los N primeros documentos, de los cuales, R han sido relevantes, calcularíamo cuántos documentos recuperados tiene t_i presente (n_i) y cuántos son relevantes (r_i).

$$p_i = \frac{r_i}{R},$$

$$q_i = \frac{n_i - r_i}{N - R}$$

$$c_i = \log \frac{\frac{r_i}{R} \cdot \frac{N - R - n_i + r_i}{N - R}}{\frac{n_i - r_i}{N - R} \cdot \frac{R - r_i}{R}} = \log \frac{r_i \cdot (N - R - n_i + r_i)}{(n_i - r_i) \cdot (R - r_i)} =$$

$$= \log \frac{\frac{r_i}{R - r_i}}{\frac{n_i - r_i}{N - R - n_i + r_i}}.$$

Para evitar que c_i tomen valores infinitos cuando R y ir toman valores bajos, c_i se suele definir como sigue:

$$c_i = \log \frac{\frac{r_i + 0{,}5}{R - r_i + 0{,}5}}{\frac{n_i - r_i + 0{,}5}{N - R - n_i + r_i + 0{,}5}}$$