FACULTAD DE CIENCIAS DE LA DOCUMENTACIÓN

MODELOS CLÁSICOS DE RECUPERACIÓN DE LA INFORMACIÓN

CUADERNO DE TRABAJO

Nº 10

Profesor

Juan Antonio Martínez Comeche

Colección Cuadernos de Trabajo, nº 10
Grado en Información y Documentación
Coordinador del Título: José Luis Gonzalo Sánchez-Molero

Primera edición

Febrero de 2013
ISBN-10: 84-695-4654-6
ISBN-13: 978-84-695-4654-3
Depósito legal: M-7202-2013
Edita: Facultad de Ciencias de la Documentación
Universidad Complutense de Madrid
C/ Santísima Trinidad, nº 37
28010 MADRID

Documento editado para ser impreso a doble cara.

Índice

Introducción

Este cuaderno de prácticas recopila ejercicios de apoyo para la asignatura “Búsqueda y Recuperación de información”, asignatura obligatoria del Grado en Información y Documentación. Uno de los objetivos principales de dicha asignatura consiste en el conocimiento de los denominados Modelos clásicos de Recuperación de información (RI). Este cuaderno aborda específicamente dicho objetivo, resumiendo las principales características de cada uno de estos modelos, junto con ejercicios que ayuden a su correcta asimilación.

Existen muchos modelos de RI, pero los denominados clásicos son tres:

- Modelo booleano.
- Modelo probabilístico.
- Modelo vectorial.

En consecuencia, en este cuaderno se encontrarán ejercicios relativos al modelo booleano, al modelo probabilístico y al modelo vectorial de Recuperación de información.

Todo modelo de Recuperación de información comporta esencialmente un método para hallar los documentos relevantes existentes en una colección en relación a cada necesidad informativa de los usuarios expresada mediante una consulta al Sistema de Recuperación de información (SRI).

Para poder llevar a cabo esta tarea, cada modelo necesita desarrollar los siguientes aspectos básicos:

- Una representación para los documentos de la colección.
- Una representación de las consultas.
- Un algoritmo o proceso que permita discernir los documentos relevantes de los irrelevantes en relación a cada consulta, y su posible ordenación.

Conforme a este esquema, cada grupo de ejercicios sobre el modelo correspondiente viene precedido por una pequeña introducción teórica sobre las tres dimensiones básicas que describen cualquier modelo de recuperación:

a) El modo de representación de los documentos.

b) El modo de representación de las consultas.

c) El modo en que se halla la respuesta del sistema ante una consulta concreta del usuario.

Los ejercicios posteriores abordan cada uno de estos aspectos en número suficiente para alcanzar una comprensión del modelo correspondiente.

MODELO BOOLEANO

A la hora de representar los documentos de la colección, el modelo booleano es un modelo binario, pues solamente considera dos posibilidades en relación a cada término:

- Valor 0 si el término está ausente en el documento
- Valor 1 si el término está presente en el documento, independientemente de su frecuencia en dicho documento.

Imaginemos que los términos presentes en todos los documentos de la colección son los siguientes: pizza, tomate, queso, bacon, peperoni, anchoas. En el documento D1="esta pizza es de tomate, bacon y anchoas", eliminadas las palabras consideradas no importantes, aparecen únicamente los términos: pizza, tomate, bacon, anchoas. Por lo tanto, estos términos tendrán valor 1 en el documento D1. El resto de términos (queso, peperoni) tendrá valor 0 en el documento D1. Dado que los términos tienen asignado un orden determinado (pizza es el término 1; tomate es el término 2; queso es el término 3; bacon es el término 4; peperoni es el término 5; y anchoas es el término 6), la representación del documento D1 conforme a este índice de términos y este orden es:

D1={1, 1, 0, 1, 0, 1}

En el modelo booleano, una consulta Q está compuesta de términos de indización unidos mediante 3 conectivas: AND, OR y NOT, pues son suficientes para expresar cualquier combinación lógica:

- "t1 AND t2" (dos términos unidos por la conectiva AND) expresa que se desean los documentos en los que aparezcan simultáneamente los términos t1 y t2.
- "t1 OR t2" (dos términos unidos por la conectiva OR) expresa que se desean los documentos en los que aparezcan o solamente el término t1 o solamente el término t2 o ambos términos simultáneamente.
- "NOT t1" (un término precedido por la conectiva NOT) expresa que se desean los

documentos en los que no aparezca el término t1.

Por último, el proceso de recuperación consiste en los siguientes dos pasos:

- En primer lugar, sustituir cada término presente en la consulta por el conjunto de documentos en los que aparece dicho término.
- A continuación, realizar las operaciones de conjuntos correspondientes a las conectivas lógicas con los conjuntos obtenidos previamente:
  - "t1 AND t2": se efectúa la INTERSECCIÓN de los conjuntos de documentos correspondientes a t1 y t2.
  - "t1 OR t2": se efectúa la UNIÓN de los conjuntos de documentos correspondientes a t1 y t2.
  - "NOT t1": se efectúa la DIFERENCIA entre el conjunto de todos los documentos de la colección y el conjunto de documentos correspondiente a t1.

Ejercicios modelo booleano

Para todos los ejercicios de esta sesión consideraremos un SRI basado en el modelo booleano básico cuyo fichero diccionario es el siguiente:

Fichero diccionario

Término	Nº total docs.	Frec. Abs.	Lista	Puntero
cafetería	3	3	1/1, 2/1, 3/1	3
caro	1	7	1/7	4
mejicano	2	6	1/3, 2/3	2
restaurante	4	13	1/4, 2/1, 3/2, 4/6	1

Ejercicio 1

Formule la consulta que debería emplear un usuario que desea localizar restaurantes mejicanos o cafeterías baratas. ¿Cuál sería la respuesta del sistema?

NOTA: Considere “baratas” equivalente a “no caras”.

Solución

La consulta booleana sería:

(restaurante AND mejicano) OR (cafetería AND (NOT caro))

Del fichero diccionario obtenemos las siguientes listas de los documentos que contienen cada uno de los términos del sistema:

U = {1, 2, 3, 4 }

Restaurante = { 1, 2, 3, 4 }

Mejicano = { 1, 2 }

Cafetería = { 1, 2, 3 }

Caro = {1}

Por lo que:

A = (restaurante AND mejicano) = {1,2,3,4} ∩ {1,2} = {1,2}

B = NOT caro = U – {1} = {1,2,3,4}- {1} = {2,3,4}

C = cafetería AND (NOT caro) = {1,2,3} ∩ {2,3,4} = {2,3}

= {1,2,3} – {1} = {2,3}

[N.B.: dado que se buscan los documentos que contengan "cafetería" pero que simultáneamente NO contengan "caro", se puede proceder eliminando de la lista correspondiente a "cafetería" los documentos de la lista correspondiente a "caro"]

Finalmente:

A OR C = {1,2} U {2,3} = {1,2,3}

La respuesta del sistema serían, pues, los documentos 1, 2 y 3 = {1,2,3} sin poder imponer ningún orden entre ellos; esto es, todos satisfacen en la misma medida las condiciones impuestas en la fórmula booleana.

Ejercicio 2

Formule la consulta correspondiente a la siguiente necesidad informativa: “restaurantes o cafeterías que sean mejicanos o baratos”. Halle la respuesta del sistema. ¿Podría expresar la misma consulta mediante una fórmula booleana con el esquema:

(term AND [NOT] term) OR (term AND [NOT] term) OR …. ?

NOTA: el NOT es optativo.

NOTA: Considere “baratos” equivalente a “no caros”

Compruebe que la respuesta del sistema sería la misma.

Solución

(restaurante OR cafeteria) AND (mejicano OR (NOT caro))

{1,2,3,4}-{1} = {2,3,4}

{1,2} OR {2,3,4} = {1,2,3,4}

{1,2,3,4} OR {1,2,3} = {1,2,3,4}

{1,2,3,4} AND {1,2,3,4} = {1,2,3,4}

RESPUESTA DEL SISTEMA = {1,2,3,4} [sin orden alguno]

(rest AND mej)OR(rest AND(NOT caro))OR(caf AND mej)OR(caf AND(NOT caro))

{1,2} OR {2,3,4} OR {1,2} OR {2,3}

{1,2,3,4}

RESPUESTA DEL SISTEMA = {1,2,3,4} [sin orden alguno]

Ejercicio 3

Un usuario desea los documentos que contengan los términos “cafetería” o “restaurante”, pero NO los dos términos simultáneamente, y que también incluyan el término “caro”. ¿Cómo formularía la consulta? ¿Qué respuesta daría el sistema?

Solución

Existen varias maneras de expresar esa necesidad informativa. Las fórmulas más habituales son las siguientes (todas ellas equivalentes entre sí):

(cafeteria OR restaurante) AND (NOT (cafeteria AND restaurante)) AND caro

[(caf AND caro) OR (rest AND caro)] AND [NOT (caf AND rest)]

[caf AND (NOT rest) AND caro] OR [rest AND (NOT caf) AND caro]

{[caf AND (NOT rest)] OR [rest AND (NOT caf)]} AND caro

Lógicamente, la respuesta del sistema ha de ser la misma sea cual sea la fórmula empleada. Nosotros utilizaremos aquí la primera:

{1,2,3,4} AND (NOT {1,2,3}) AND {1}

{1,2,3,4} ∩ {4} ∩ {1}

{4} ∩ {1}

ø

LA RESPUESTA DEL SISTEMA SERÍA EL CONJUNTO VACÍO; ESTO ES, NO HAY NINGÚN DOCUMENTO EN LA COLECCIÓN QUE VERIFIQUE LAS CONDICIONES IMPUESTAS EN LA CONSULTA.

Ejercicio 4

¿Cómo preguntaría por “restaurantes mejicanos, preferiblemente caros”? ¿Cuál sería la respuesta del sistema? ¿Y si preguntásemos por “restaurantes mejicanos”? ¿Qué ventaja aporta la primera expresión booleana frente a la segunda?

Solución

La primera consulta booleana sería la siguiente:

(restaurante AND mejicano AND caro) OR (restaurante AND mejicano)

$(\{1,2,3,4\} \cap \{1,2\} \cap \{1\}) \cup (\{1,2,3,4\} \cap \{1,2\})$

$(\{1,2,3,4\} \cap \{1\}) \cup (\{1,2\})$

$\{1\} \cup \{1,2\}$

$\{1,2\}$

La segunda consulta booleana sería la siguiente:

(restaurante AND mejicano)

$\{1,2,3,4\} \cap \{1,2\}$

{1,2}

LA RESPUESTA SERÍA LA MISMA.

La ventaja de la primera expresión frente a la segunda radica en que el sistema podría memorizar los resultados parciales de la búsqueda -en nuestro caso, el conjunto {1} correspondiente a (restaurante AND mejicano AND caro)-.

Ahora bien, al tratarse la consulta en su conjunto de un OR lógico entre dicho conjunto {1} y otro más amplio {1,2}, el sistema podría **ORDENAR LA RESPUESTA** "RAZONANDO" QUE EL RESULTADO PARCIAL MÁS ESPECÍFICO –{1} en este caso- SERÁ MÁS RELEVANTE PARA EL USUARIO PUES VERIFICA LA CONDICIÓN MÁS RESTRICTIVA DENTRO DE LA CONSULTA, "AMPLIADA" POSTERIORMENTE MEDIANTE UN "OR" AL RESTO DE LOS DOCUMENTOS –el {2} del conjunto {1,2} obtenido finalmente-.

En consecuencia, el sistema podría imponer un orden en la respuesta (en primer lugar el documento {1}; en segundo lugar el documento {2}) superando así uno de los inconvenientes del modelo booleano básico.

Este es uno de los procedimientos de los que pueden valerse los denominados modelos booleanos extendidos.

Ejercicio 5

Un usuario desea localizar información sobre dos cualesquiera términos (siempre que aparezcan simultáneamente) de los cuatro iniciales (restaurante, mejicano, cafetería y caro). Formule la consulta y halle la respuesta del sistema.

Solución

(rest AND mej) OR (rest AND caf) OR (rest AND caro) OR (mej AND caf) OR (mej AND caro) OR (caf AND caro)

{1,2} U {1,2,3} U {1} U {1,2} U {1} U {1}

{1,2,3}

Algunos sistemas booleanos permiten que el usuario pueda introducir una lista de términos, quedando satisfecho con que un número de ellos aparezca simultáneamente en los documentos resultantes. La interfaz suele emplear en estos casos la conectiva OF, de manera que la consulta del ejercicio se introduciría:

2 OF (restaurante, mejicano, cafetería, caro)

El sistema se ocupa de "traducirlo" a la fórmula booleana correspondiente que hemos desarrollado inicialmente.

Ejercicio 6

Determine las combinaciones de “restaurante”, “cafetería” y “mejicano”(ejemplo hipotético: “restaurante” presente; “cafetería” ausente; “mejicano” ausente) que figurarán en los documentos recuperados por el sistema para las siguientes consultas:

restaurante OR (cafeteria AND mejicano)

(restaurante OR cafeteria) AND mejicano

Solución

PRIMERA FÓRMULA: restaurante OR (cafeteria AND mejicano)

La fórmula corresponde al esquema: A **OR** B.

En consecuencia, las combinaciones posibles que verifican esta fórmula booleana son las siguientes:

1ª.- A se cumple ; B se cumple.

2ª.- A se cumple ; B no se cumple.

3ª.- A no se cumple ; B se cumple.

En nuestro caso, las combinaciones posibles son:

1ª.- restaurante SE CUMPLE ; (cafeteria AND mejicano) SE CUMPLE

2ª.- restaurante SE CUMPLE ; (cafeteria AND mejicano) NO SE CUMPLE

3ª.- restaurante NO SE CUMPLE ; (cafeteria AND mejicano) SE CUMPLE

Por otra parte:

a) restaurante se cumple cuando aparece en el documento.
b) restaurante no se cumple cuando no aparece en el documento.
c) Para que se verifique la fórmula "A AND B" es preciso que simultáneamente A se cumpla y B se cumpla.

En consecuencia:

1ª.- [restaurante SE CUMPLE ; (cafeteria AND mejicano) SE CUMPLE] implica:

1.1. restaurante aparece; cafetería aparece; mejicano aparece

2ª.- [restaurante SE CUMPLE ; (cafeteria AND mejicano) NO SE CUMPLE] implica

restaurante aparece ; {no aparecen simultáneamente cafetería y mejicano; o bien no aparece ninguno de los dos términos}

Es decir:

2.1. restaurante aparece; cafetería aparece; mejicano no aparece

2.2. restaurante aparece; cafetería no aparece; mejicano aparece

2.3. restaurante aparece; cafetería no aparece; mejicano no aparece

3ª.- [restaurante NO SE CUMPLE ; (cafeteria AND mejicano) SE CUMPLE] implica

3.1. restaurante no aparece; cafetería aparece; mejicano aparece

Reuniendo todas las posibilidades deducidas, obtenemos que en los documentos recuperados por el sistema son posibles las siguientes combinaciones de términos:

restaurante aparece; cafetería aparece; mejicano aparece

restaurante aparece; cafetería aparece; mejicano no aparece

restaurante aparece; cafetería no aparece; mejicano aparece

restaurante aparece; cafetería no aparece; mejicano no aparece

restaurante no aparece; cafetería aparece; mejicano aparece

SEGUNDA FÓRMULA: (restaurante OR cafetería) AND mejicano

La fórmula corresponde al esquema: A **AND** B.

En consecuencia, la única combinación posible que verifica esta fórmula booleana es: A se cumple ; B se cumple.

En nuestro caso:

(restaurante OR cafeteria) SE CUMPLE; mejicano SE CUMPLE

Ello implica:

{al menos uno de los dos –restaurante, cafetería- aparece}; mejicano aparece

En relación a la fórmula (restaurante OR cafeteria), las posibilidades son:

1. restaurante aparece; cafetería aparece
2. restaurante aparece; cafetería no aparece
3. restaurante no aparece; cafetería aparece

Finalmente, como “mejicano” debe aparecer siempre, obtenemos que en los documentos recuperados por el sistema son posibles las siguientes combinaciones de términos:

restaurante aparece; cafetería aparece; mejicano aparece

restaurante aparece; cafetería no aparece; mejicano aparece

restaurante no aparece; cafetería aparece; mejicano aparece

Ejercicio 7

Demuestre gráficamente que el conjunto de documentos recuperados con la consulta

(restaurante AND mejicano) OR (restaurante AND caro)

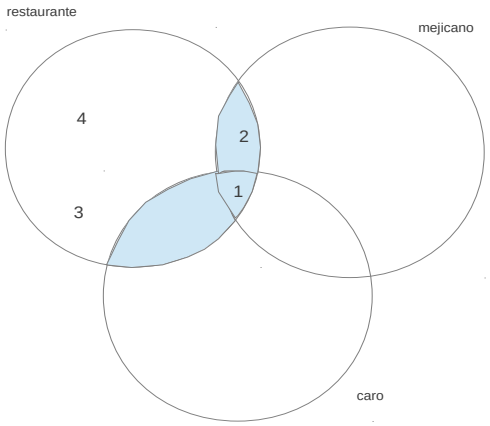
es diferente del conjunto de documentos recuperados con la consulta

(restaurante OR mejicano) AND (restaurante OR caro)

NOTA: Esto es, demostrar gráficamente que las fórmulas no son equivalentes.

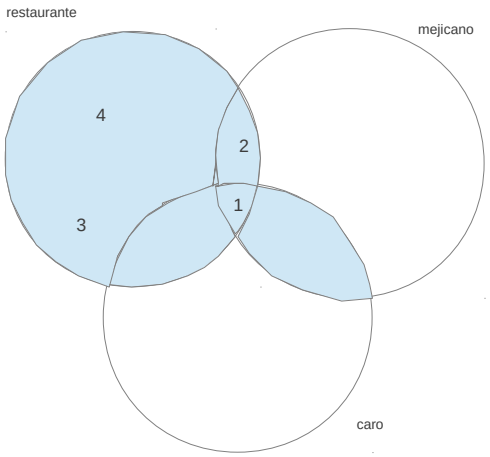
Solución

(restaurante AND mejicano) OR (restaurante AND caro)



RESPUESTA DEL SISTEMA = {1,2} [sin orden alguno]

(restaurante OR mejicano) AND (restaurante OR caro)



RESPUESTA DEL SISTEMA = {1,2,3,4} [sin orden alguno]

MODELO PROBABILÍSTICO

A la hora de representar los documentos de la colección, el modelo probabilístico es también un modelo binario, pues solamente considera dos posibilidades en relación a cada término:

- Valor 0 si el término está ausente en el documento.
- Valor 1 si el término está presente en el documento, independientemente de su frecuencia en dicho documento.

En consecuencia, los documentos de la colección son representados mediante una serie ordenada de ceros y unos, como sucedía en el modelo booleano.

En el modelo probabilístico, una consulta Q está compuesta de la enumeración de los términos de indización que deseamos encontrar en los documentos de la respuesta. Por tanto, una consulta Q tiene la misma representación que un documento de la colección; esto es, una serie ordenada de ceros y unos. Los unos corresponderán a los términos de la colección que deseamos hallar en los documentos de la respuesta.

Por último, el proceso de recuperación consiste en los siguientes pasos:

- Cálculo de los coeficientes iniciales (o valor del estado de recuperación; RSV o Retrieval Status Value en inglés) correspondientes a cada término de la consulta, mediante la fórmula:

$$c_i = \log\left(\frac{N - n_i}{n_i}\right)$$

Siendo N = número total de documentos de la colección; n_i = número de documentos de la colección que contienen el término i.

- Cálculo de la similitud entre cada documento de la colección y la consulta, mediante la suma de los coeficientes iniciales correspondientes a los términos

presentes simultáneamente en el documento y en la consulta:

$SIM(D_i, Q) = \Sigma c_i$

- Mostrar al usuario los documentos ordenados en orden decreciente de su similitud con la consulta.
- Nuevo cálculo de los coeficientes conforme a la clasificación en relevantes e irrelevantes efectuada por el usuario.
- Nuevo cálculo de similitud entre cada documento de la colección y la consulta.
- Mostrar al usuario los documentos ordenados en orden decreciente de su similitud con la consulta.

Ejercicios modelo probabilístico

Ejercicio 1

Considere un SRI basado en el modelo probabilístico básico cuyo fichero diccionario es aproximadamente el mismo que el empleado en el modelo booleano de la sesión anterior:

Fichero diccionario

Término	Nº total docs.	Frec. Abs.	Lista	Puntero
Barato	1	1	5/1	5
Cafetería	3	3	1/1, 2/1, 3/1	3
Caro	1	7	1/7	4
Mejicano	2	6	1/3, 2/3	2
Restaurante	4	13	1/4, 2/1, 3/2, 4/6	1

Hallar la descripción de los cinco documentos de la colección. Expresar la consulta sobre "cafeterías mejicanas preferiblemente caras" y hallar la respuesta inicial del sistema. ¿Figura el documento 4 en la respuesta? ¿En algún orden específico? ¿Figuraría en dicho orden si se tratase de un SRI basado en el modelo booleano?

Solución

A) Tratándose de un modelo probabilístico, el SRI emplea internamente un método binario de representación de los documentos de la colección; esto es, únicamente utiliza 1 (para indicar la presencia de un término en un documento) y 0 (para indicar la ausencia de un término en un documento). De este modo, la descripción de los cinco documentos de la colección sería:

D1 = < 1, 1, 1, 1, 0 >

D2 = < 1, 1, 1, 0, 0 >

D3 = < 1, 0, 1, 0, 0 >

D4 = < 1, 0, 0, 0, 0 >

D5 = < 0, 0, 0, 0, 1 >

Teniendo en cuenta que:

* El primer nº corresponde a restaurante

* El segundo nº corresponde a mejicano

* El tercer nº corresponde a cafetería

* El cuarto nº corresponde a caro

* El quinto nº corresponde a barato

B-1) En un modelo probabilístico básico, el SRI únicamente necesita que en la consulta aparezcan aquellos términos empleados por el usuario y que simultáneamente figuren en el fichero inverso del sistema. El resto es directamente desechado. En consecuencia, la consulta quedaría formulada de la siguiente manera en el sistema:

Q = cafeterías mejicanas caras

cuya representación interna sería la siguiente:

Q = < 0, 1, 1, 1, 0 >

Nótese que no figura ninguna conectiva entre los términos. Ello es debido a que el modelo probabilístico no considera la lógica booleana ni en la formulación de la consulta ni en el algoritmo de búsqueda de la respuesta.

B-2) Para hallar la respuesta del sistema, efectuamos la siguiente hipótesis inicial:

p(cafeterías/R) = 0'5 $$p(cafeterías / \overline{R}) = \frac{3}{5}$$

p(mejicanas/R) = 0’5 $$p(mejicanas / \overline{R}) = \frac{2}{5}$$

p(caro/R) = 0’5 $$p(caro / \overline{R}) = \frac{1}{5}$$

Por otra parte, los cálculos de la similitud entre los documentos de la colección y la consulta Q parten de las descripciones o representaciones internas de los documentos y de la consulta que ya hemos hallado anteriormente:

Q = < 0, 1, 1, 1, 0 >

D1 = < 1, 1, 1, 1, 0 >

D2 = < 1, 1, 1, 0, 0 >

D3 = < 1, 0, 1, 0, 0 >

D4 = < 1, 0, 0, 0, 0 >

D5 = < 0, 0, 0, 0, 1 >

Aplicando la fórmula correspondiente (vid el apartado “Modelo probabilístico” más arriba), tenemos:

sim(D1,Q) = c(mej) + c(cafet) + c(caro)

sim(D2,Q) = c(mej) + c(cafet)

sim(D3,Q) = c(cafet)

sim(D4,Q) = 0

sim(D5,Q) = 0

Falta calcular, pues, los valores de los coeficientes correspondientes a los términos "mejicanas", "cafeterías" y "caro". Para ello nos basaremos en la hipótesis inicial expuesta anteriormente y en las fórmulas que figuran en el apartado "Modelo probabilístico". Los resultados son los siguientes:

$$c(mejicanas) = \log(\frac{0'5}{1-0'5}) + \log(\frac{1-\frac{2}{5}}{\frac{2}{5}}) = \log\frac{0'5}{0'5} + \log\frac{\frac{5-2}{5}}{\frac{2}{5}} = 0 + \log\frac{\frac{3}{5}}{\frac{2}{5}} = \log\frac{3}{2}$$

$$c(cafeterías) = \log(\frac{0'5}{1-0'5}) + \log(\frac{1-\frac{3}{5}}{\frac{3}{5}}) = 0 + \log\frac{\frac{2}{5}}{\frac{3}{5}} = \log\frac{2}{3}$$

$$c(caro) = \log(\frac{0'5}{1-0'5}) + \log(\frac{1-\frac{1}{5}}{\frac{1}{5}}) = \log\frac{\frac{4}{5}}{\frac{1}{5}} = \log\frac{4}{1} =$$

Con estos valores sustituimos en las fórmulas de las similitudes y obtenemos:

sim(D1,Q) = log3/2 + log2/3 + log 4 = log 4 = 0'602

sim(D2,Q) = log3/2 + log2/3 = 0

sim(D3,Q) = log 2/3 = -0'176

sim(D4,Q) = 0

sim(D5,Q) = 0

En resumen, la respuesta del sistema sería, **EN ESTE ORDEN**:

D1

D2, D4, D5 [en cualquier orden]

D3

C) Al tratarse de un sistema probabilístico, en principio **todos los documentos de la colección pueden ser ordenados en la respuesta a cualquier consulta** realizada por el usuario, siempre en orden decreciente de probabilidad de relevancia en relación a dicha consulta.

En consecuencia, el documento 4 figura en la respuesta.

D) Al tratarse de un sistema probabilístico, cualquier documento figura en la respuesta **en un orden preciso de probabilidad de relevancia en relación a dicha consulta**. En este caso, el documento 4 aparecerá en el segundo lugar (junto a otros dos documentos) de los cinco posibles.

E) Si el SRI estuviera basado en un modelo booleano, en ningún caso el documento 4 podría figurar en dicho lugar, pues los modelos booleanos básicos **no pueden efectuar la ordenación de los documentos en la respuesta dada al usuario.** De hecho, el documento D4, en un modelo booleano, figuraría o no figuraría en la respuesta, dependiendo de si fuese relevante o no en relación a la consulta. En nuestro caso, si se tratase de un modelo booleano, la consulta sería equivalente a (vid. ejercicios modelo booleano):

Q = cafeterías AND mejicanas AND caras

$$R = (\{1,2,3\} \cap \{1,2\}) \cap \{1\} = \{1,2\} \cap \{1\} = \{1\}$$

Como puede observarse, solamente el documento D1 figuraría en la respuesta si se tratase de un modelo booleano, de manera que el documento D4 ni siquiera aparecería en la respuesta. Al contrario, si el SRI se basa en un modelo probabilístico, el documento D4 aparece en la respuesta, aunque en segundo lugar.

Ejercicio 2

Sea el mismo sistema anterior, basado en el modelo probabilístico básico. Expresar la consulta sobre “cafeterías caras o cafeterías mejicanas”. ¿La respuesta inicial del sistema sería idéntica a la obtenida en el ejercicio anterior? Una vez mostrada la respuesta inicial, se solicita al usuario que elija los documentos que considera relevantes. El usuario elige los documentos D1, D2 y D4. Hallar la nueva respuesta del sistema a dicha consulta. ¿Es distinta de la primera?

Solución

A) Al tratarse de un sistema basado en el modelo probabilístico, únicamente podemos expresar qué términos deseamos que estén presentes en la consulta y cuáles no queremos que figuren en ella. Por tanto, la consulta sobre “cafeterías caras o cafeterías mejicanas” incluiría los términos “cafeterías”, “mejicanas” y “caras”. Como dijimos anteriormente, el resto es directamente desechado. En consecuencia, la consulta quedaría formulada de la siguiente manera en el sistema:

$$Q = \text{cafeterías mejicanas caras}$$

cuya representación interna sería la siguiente:

$$Q = < 0, 1, 1, 1, 0 >$$

En consecuencia, tanto la consulta como la respuesta inicial del sistema serían idénticas a las obtenidas en el ejercicio anterior.

B) Una vez que el usuario ha elegido los documentos D1, D2 y D4 como relevantes, para hallar la nueva respuesta del sistema partimos de los dos conjuntos siguientes:

Conjunto de documentos relevantes: D1, D2, D4.

Conjunto de documentos irrelevantes: D3, D5 [el resto de la colección]

Ahora las probabilidades de ocurrencia de los términos son las siguientes:

p(cafeterías/R) = 2/3 $p(cafeterías / \overline{R}) = 0{,}5$

p(mejicanas/R) = 2/3 $p(mejicanas / \overline{R}) = 0$

p(caro/R) = 1/3 $p(caro / \overline{R}) = 0$

Por lo que los nuevos valores de los coeficientes son:

$$c(mejicanas) = \log\left(\frac{\frac{2}{3}}{1-\frac{2}{3}}\right) + \log\left(\frac{1-0}{0}\right) = \log 2 + \log\frac{1}{0} = \log 2 + \log\infty = +\infty$$

$$c(cafeterías) = \log\left(\frac{\frac{2}{3}}{1-\frac{2}{3}}\right) + \log\left(\frac{1-0'5}{0'5}\right) = \log 2 + \log 1 = \log 2$$

$$c(caro) = \log\left(\frac{\frac{1}{3}}{1-\frac{1}{3}}\right) + \log\left(\frac{1-0}{0}\right) = \log\frac{1}{2} + \log\frac{1}{0} = -\log 2 + \log\infty = +\infty$$

Conforme a la lógica del modelo probabilístico, los términos “mejicanas” y “caro” aparecen con distinta distribución en los documentos relevantes, pero NO APARECEN NI UNA

SOLA VEZ en los documentos irrelevantes (ese es el juicio emitido por el usuario), de manera que COMO ÚNICAMENTE APARECEN EN LOS DOCUMENTOS RELEVANTES, los coeficientes correspondientes a “mejicanas” y “caro” serían lo más positivo posible desde el punto de vista matemático, esto es,  (más infinito).

Si mantenemos el valor del coeficiente de los términos “mejicanas” y “caro” no podríamos ordenar los documentos de la colección en cuya similitud aparecieran ambos valores (como sucede en nuestro caso). Por tanto, para poder efectuar comparaciones, consideraremos que su valor es N, siendo N extremadamente grande pero finito. De esta forma, tenemos:

$$sim(D1,Q) = c(mej) + c(cafet) + c(caro)$$

$$sim(D2,Q) = c(mej) + c(cafet)$$

$$sim(D3,Q) = c(cafet)$$

$$sim(D4,Q) = 0$$

$$sim(D5,Q) = 0$$

Sustituyendo:

$$sim(D1,Q) = (log2 + N) + (log2) + (-log2 + N)=2N + log2$$

$$sim(D2,Q) = (log2 + N) + log2 = N + 2log2$$

$$sim(D3,Q) = log2$$

$$sim(D4,Q) = sim(D5,Q) = 0$$

En resumen, la nueva respuesta del sistema sería, **EN ESTE ORDEN**:

D1

D2

D3

D4, D5 [en cualquier orden]

Conviene fijarse en dos aspectos de la respuesta:

α) En primer lugar, el documento D2 asciende a ocupar la segunda plaza en solitario, consecuencia de dos factores que actúan simultáneamente: D2 posee los términos "mejicanas" y "cafeterías" que aparecen en la consulta; y ha sido señalado como relevante por el usuario junto a D1, por lo que "mejicanas" únicamente aparece en documentos relevantes, poseyendo en consecuencia una altísima valoración como descriptor del conjunto buscado por el usuario (+infinito).

β) En segundo lugar, el documento D4 desciende a ocupar la cuarta y última plaza, a pesar de haber sido señalado como relevante por el usuario. En ese comportamiento del sistema influye un único factor: DICHO DOCUMENTO NO POSEE NINGUNO DE LOS TÉRMINOS DE LA CONSULTA, POR LO QUE SEGUIRÁ PERMANECIENDO NECESARIAMENTE EN LAS POSICIONES MÁS BAJAS DE LA LISTA (en el cálculo de su similitud con la consulta no podrá haber coeficientes que contribuyan cuantitativamente en su apoyo).

Ejercicio 3

Sea un SRI basado en el modelo probabilístico básico cuyo fichero diccionario es el mismo que el empleado en el modelo booleano de la sesión anterior:

Fichero diccionario

Término	Nº total docs.	Frec. Abs.	Lista	Puntero
Cafetería	3	3	1/1, 2/1, 3/1	3
Caro	1	7	1/7	4
Mejicano	2	6	1/3, 2/3	2
Restaurante	4	13	1/4, 2/1, 3/2, 4/6	1

Halle la respuesta inicial del sistema ante la consulta sobre “restaurantes caros”. Compare la respuesta del sistema si el SRI emplea el modelo booleano básico o el modelo probabilístico básico.

Solución

A) Tanto si el SRI se basa en el modelo probabilístico como si se basa en el modelo booleano, la descripción de los documentos es binaria:

D1 = < 1, 1, 1, 1 >

D2 = < 1, 1, 1, 0 >

D3 = < 1, 0, 1, 0 >

D4 = < 1, 0, 0, 0 >

La consulta, en cambio, varía de expresión si el SRI se basa en el modelo booleano o en el modelo probabilístico. En el modelo booleano sería:

Q = restaurantes AND caros

Mientras que en el modelo probabilístico sería:

Q = < 1, 0, 0, 1 >

Para hallar la respuesta inicial del sistema en este último caso, partimos de la siguiente hipótesis inicial:

p(restaurantes/R) = 0'5 $p(restaurantes/\overline{R}) = \frac{4}{4}$

p(caro/R) = 0'5 $p(caro/\overline{R}) = \frac{1}{4}$

Por otra parte, los cálculos de la similitud entre los documentos de la colección y la consulta Q son los siguientes:

Q = < 1, 0, 0, 1 >

D1 = < 1, 1, 1, 1 >

D2 = < 1, 1, 1, 0 >

D3 = < 1, 0, 1, 0 >

D4 = < 1, 0, 0, 0 >

sim(D1,Q) = c(rest) + c(caro)

sim(D2,Q) = c(rest)

sim(D3,Q) = c(rest)

sim(D4,Q) = c(rest)

A su vez, los valores de los coeficientes correspondientes a los términos “restaurante” y “caro” son:

$$c(restaurantes) = \log(\frac{0'5}{1-0'5}) + \log(\frac{1-1}{1}) = 0 + \log\frac{0}{1} = \log 0 = -\infty$$

$$c(caro) = \log(\frac{0'5}{1-0'5}) + \log(\frac{1-\frac{1}{4}}{\frac{1}{4}}) = \log\frac{\frac{3}{4}}{\frac{1}{4}} = \log\frac{3}{1} =$$

En efecto, conforme a la lógica del modelo probabilístico, el término “restaurantes” aparece en todos los documentos irrelevantes (esa es la hipótesis inicial en nuestro caso), de manera que el coeficiente correspondiente a restaurantes sería el más negativo posible desde el punto de vista matemático, esto es, (menos infinito).

Si mantenemos el valor del coeficiente del término “restaurantes”, todas las similitudes valdrían , imposibilitando la ordenación de los documentos de la colección. Dado que dicho componente es común a todas las similitudes, podemos eliminar el componente

correspondiente al coeficiente del término “restaurante” en todas las similitudes de todos los documentos de la colección. De esta forma:

sim(D1,Q) = c(caro) = log 3 = 0’477

sim(D2,Q) = 0

sim(D3,Q) = 0

sim(D4,Q) = 0

En definitiva, la respuesta del sistema sería:

D1

D2, D3, D4 [en cualquier orden]

B) La respuesta del sistema, basándose en el modelo booleano, sería la siguiente:

Q = restaurantes AND caros

R = {1,2,3,4} ∩ {1} = {1}

En resumen, la respuesta del SRI hubiera sido: D1

En este caso, como puede observarse, no existe diferencia alguna entre la respuesta del sistema si se emplea un modelo booleano o un modelo probabilístico (en un modelo booleano también podríamos haber añadido los demás documentos de la colección posteriormente al documento D1 sin necesidad de efectuar ningún cálculo). En estos casos, pues, la ventaja del empleo del modelo probabilístico se diluye. Sin embargo, debe advertirse que son extremadamente extraños en la práctica.

Ejercicio 4

Sea un SRI basado en el modelo probabilístico básico cuya colección consta de 100 documentos caracterizados por 5 términos con la siguiente distribución:

- El término t_1 aparece en 10 documentos de los 100 de la colección.
- El témino t_2 aparece en 20 documentos de los 100 de la colección.
- El término t_3 aparece en 30 documentos de los 100 de la colección.
- El término t_4 aparece en 40 documentos de los 100 de la colección.
- El término t_5 aparece en 5 documentos de los 100 de la colección.

Seis de los documentos de la colección están representados de la siguiente manera:

D1 = < 0, 1, 0, 1, 0 >

D2 = < 1, 1, 0, 0, 1 >

D3 = < 0, 0, 1, 1, 1 >

D4 = < 0, 1, 1, 0, 1 >

D5 = < 1, 0, 0, 1, 0 >

D6 = < 0, 0, 1, 1, 0 >

Se efectúa al SRI la siguiente consulta: Q = < 1, 1, 0, 0, 1 >.

A) Halle la respuesta del sistema a dicha consulta, en relación a D1, D2….D6.
B) Obtenida la respuesta, se solicita al usuario que emita un juicio sobre los que considera relevantes y los que juzga irrelevantes. El usuario señala como relevantes 19 documentos en total, con la siguiente distribución:
    i. t_1 aparece en 3 documentos de los 19
    ii. t_2 aparece en 18 documentos de los 19
    iii. t_5 aparece en 2 documentos de los 19

    Se considerarán, pues, irrelevantes los demás documentos de la colección. Halle la nueva respuesta del sistema a dicha consulta, en relación a D1, D2, D3,….., D6. ¿Es distinta de la primera? Analice el comportamiento del sistema con respecto al documento D1.

Solución

A) El SRI efectúa una hipótesis inicial sobre la distribución de los términos (únicamente aquellos presentes en la consulta) en los conjuntos de documentos relevantes e irrelevantes de la colección respecto a la consulta Q

$$p(t1/R) = 0'5 \qquad p(t1/\overline{R}) = \frac{n1}{N} = \frac{10}{100} = 0'1$$

$$p(t2/R) = 0'5 \qquad p(t2/\overline{R}) = \frac{n2}{N} = \frac{20}{100} = 0'2$$

$$p(t5/R) = 0'5 \qquad p(t5/\overline{R}) = \frac{n5}{N} = \frac{5}{100} = 0'05$$

A continuación el SRI efectúa el cálculo de los coeficientes o pesos iniciales correspondientes a los términos presentes en la consulta:

$$c1 = \log\frac{0'5}{0'5} + \log\frac{1-0'1}{0'1} = \log 9$$

$$c2 = \log\frac{1-0'2}{0'2} = \log\frac{0'8}{0'2} = \log 4$$

$$c5 = \log\frac{1-0'05}{0'05} = 1'279$$

Luego el SRI calcula la similitud entre cada documento de la colección y la consulta Q mediante la suma de los coeficientes o pesos correspondientes a los términos presentes simultáneamente en el documento y en la consulta:

$sim(D1,Q) = c2 = \log 4 = 0'602$

$sim(D2,Q) = c1 + c2 + c5 = 2'835$

$sim(D3,Q) = c5 = 1'279$

$sim(D4,Q) = c2 + c5 = 1'881$

$sim(D5,Q) = c1 = 0'954$

$sim(D6,Q) = 0$

En consecuencia, el SRI mostraría, en su respuesta, los documentos ordenados en orden decreciente de su probabilidad de relevancia en relación a la consulta:

D2

D4

D3

D5

D1

D6

B) El SRI calcula de nuevo la distribución de los términos de la consulta en los conjuntos de relevantes e irrelevantes definidos por el usuario:

$$p(t1/R) = \frac{|V1|}{|V|} = \frac{3}{19} = 0'158 \qquad p(t1/\overline{R}) = \frac{n1 - |V1|}{N - |V|} = \frac{10-3}{100-19} = \frac{7}{81} = 0'086$$

$$p(t2/R) = \frac{|V2|}{|V|} = \frac{18}{19} = 0'947 \qquad p(t2/\overline{R}) = \frac{n2 - |V2|}{N - |V|} = \frac{20-18}{81} = 0'025$$

$$p(t5/R) = \frac{|V5|}{|V|} = \frac{2}{19} = 0'105 \qquad p(t5/\overline{R}) = \frac{n5 - |V5|}{N - |V|} = \frac{5-2}{81} = 0'037$$

El SRI calcula los nuevos coeficientes o pesos correspondientes a los términos presentes en la consulta, en base al juicio de relevancia emitido por el usuario:

$$c1 = \log\frac{0'158}{1-0'158} + \log\frac{1-0'086}{0'086} = -0'727 + 1'026 = 0'299$$
$$c2 = \log\frac{0'947}{1-0'947} + \log\frac{1-0'025}{0'025} = 1'252 + 1'591 = 2'843$$
$$c3 = \log\frac{0'105}{1-0'105} + \log\frac{1-0'037}{0'037} = -0'931 + 1'415 = 0'484$$

El SRI calcula de nuevo la similitud entre cada documento de la colección y la consulta Q, mediante la suma de los coeficientes o pesos correspondientes a los términos presentes simultáneamente en el documento y en la consulta:

$$sim(D1,Q) = c2 = 2'843$$
$$sim(D2,Q) = c1 + c2 + c5 = 3'626$$
$$sim(D3,Q) = c5 = 0'484$$
$$sim(D4,Q) = c2 + c5 = 3'327$$
$$sim(D5,Q) = c1 = 0'299$$
$$sim(D6,Q) = 0$$

Finalmente, el SRI mostraría en su respuesta los documentos en orden decreciente de probabilidad de relevancia en relación a la consulta:

D2

D4

D1

D3

D5

D6

Como vemos, D1 ha subido en posición relativa. Ello es debido a que D1 también contiene el término t_2, cuya presencia en los documentos relevantes (juzgados así por el usuario) es abrumadora.

MODELO VECTORIAL

El modelo vectorial representa los documentos de la colección como una serie ordenada de números reales positivos, en lugar de unos y ceros como los modelos anteriores. Con ello trata de reflejar la importancia de cada término en cada documento. Son dos los factores habitualmente considerados a la hora de imponer un peso a cada término en cada documento:

- **tf** (term frequency) representa la frecuencia del término en el documento.
- **idf** (inverse document frequency) representa el inverso del número de documentos de la colección en los que aparece el término. Es un número mayor cuanto menor es el número de documentos en los que aparece un término. Es un indicador, pues, de la especificidad del término.

La representación de una consulta Q en el modelo vectorial es la misma que un documento, esto es, una serie ordenada de números reales positivos. En el caso de las consultas, los números reflejan la importancia relativa de cada término en la necesidad informativa del usuario.

Por último, el proceso de recuperación consiste en los siguientes pasos:

- Hallar la descripción de los documentos de la colección conforme a un esquema de ponderación dado, tf.idf, por ejemplo.
- Hallar la similitud de cada documento de la colección con la consulta conforme a una función de similitud dada, el coseno del ángulo que forman, por ejemplo.
- Mostrar al usuario los documentos ordenados en orden decreciente de su similitud con la consulta.

Ejercicios modelo vectorial

Ejercicio 1

Considere un SRI basado en el modelo vectorial básico cuyo fichero diccionario es el siguiente:

Fichero diccionario

Término	Nº total docs.	Frec. Abs.	Lista	Puntero
Bajo	3	9	1/1, 2/3, 3/5	1
Calma	2	3	2/2, 3/1	2
Luna	1	8	2/8	3

Hallar la descripción de los tres documentos de la colección, empleando ponderación tf.idf. Compárelo con una ponderación en que solo se tuviese en cuenta la frecuencia de aparición de cada término en cada documento y explique las coincidencias y divergencias observadas.

Solución

A) Tratándose de un modelo vectorial con ponderación tf.idf, el SRI representa internamente los documentos de la colección multiplicando dos factores: la frecuencia de los términos en los documentos (factor tf) y el valor discriminatorio de los términos, calculado aquí mediante la frecuencia inversa de los documentos (factor idf). De este modo, la descripción de los documentos en el modelo vectorial ya no es binaria (como en los modelos booleano y probabilístico) , sino mediante números reales. Empezaremos considerando el factor tf, esto es, la frecuencia de los términos en los documentos. Observando los datos del

fichero diccionario, la descripción de los tres documentos de la colección empleando exclusivamente este primer factor sería:

D1 = < 1, 0, 0 >

D2 = < 3, 2, 8 >

D3 = < 5, 1, 0 >

Calcularemos a continuación los factores IDF correspondientes a cada uno de los tres términos empleados en la descripción de los documentos de la colección:

$$IDF1 = \log(\frac{N}{DOCFREQ1}) + 1 = \log(\frac{3}{3}) + 1 = \log 1 + 1 = 0 + 1 = 1$$

$$IDF2 = \log(\frac{N}{DOCFREQ2}) + 1 = \log(\frac{3}{2}) + 1 = 1'176$$

$$IDF3 = \log(\frac{N}{DOCFREQ3}) + 1 = \log(\frac{3}{1}) + 1 = \log 3 + 1 = 1'477$$

Ahora debemos multiplicar los valores del factor tf por los correspondientes al factor idf. Para ello, la primera columna (donde están situados los respectivos valores del término 1) se multiplicará por el factor IDF_1; la segunda columna (donde están situados los respectivos valores del término 2) se multiplicará por el factor IDF_2; finalmente, la tercera columna (donde están situados los valores del término 3) se multiplicará por el factor IDF_3.

En consecuencia, la descripción de los tres documentos de la colección, empleando ponderación tf.idf, será:

D1 = < 1, 0 , 0 >

D2 = < 3, 2’352, 11’816 >

D3 = < 5, 1’176, 0 >

B) Para comparar las dos representaciones, ambas obtenidas previamente, las dispondremos de manera que a la izquierda figure la ponderación tf y a su derecha la ponderación tf.idf:

D1 = < 1, 0, 0 >	D1 = < 1, 0 , 0 >
D2 = < 3, 2, 8 >	D2 = < 3, 2’352, 11’816 >
D3 = < 5, 1, 0 >	D3 = < 5, 1’176, 0 >

De la comparación de las dos posibles descripciones se obtienen las siguientes conclusiones:

- El documento 1 presenta la misma representación porque:
  - Los términos que no aparecen en un documento siempre tienen peso cero (0).
  - Como el término 1 aparece en todos los documentos de la colección, su factor IDF es el mínimo posible, esto es, uno (1).
- El documento 2 se caracteriza por un aumento importante en el peso del término 3, debido a que su factor IDF es el máximo posible porque solo aparece en este documento 2.
- Por último, el documento 3 mantiene 2 valores y aumenta un tercero:

- Mantine 2 valores porque o es un cero (0) o el término tiene un factor IDF igual a uno (1) (el mínimo posible, pues aparece en todos los documentos de la colección);
- Aumenta 1 valor correspondiente al término 2. Ello es debido a que su IDF > 1 (porque aparece en dos de los tres documentos de la colección).

Ejercicio 2

Sea un SRI basado en el modelo vectorial básico, compuesto por tres documentos cuya representación es la siguiente:

D1= < 1’477 , 0 , 1’301 >

D2= < 0 , 3’556 , 10’408 >

D3= < 7’386 , 1’778 , 0 >

Un usuario del sistema formaliza su necesidad informativa mediante la siguiente consulta:

Q= < 2’954 , 1’778 , 3’903 >

Si el sistema emplea el coeficiente del coseno para el cálculo de similitudes, ¿cuál sería el orden de presentación de los documentos como respuesta a dicha consulta del usuario? ¿Se obtendría el mismo resultado si el sistema emplease el coeficiente de Jaccard?

Solución

A) Si el SRI emplea el coeficiente del coseno para el cálculo de las similitudes, tendremos:

$$SIM(D1,Q) = \frac{\sum_{k=1}^{t}(TERMik.TERMjk)}{\sqrt{\sum_{k=1}^{t}(TERMik)^2.\sum_{k=1}^{t}(TERMjk)^2}} = \frac{2'954.1'477 + 3'903.1'301}{\sqrt{(1'477^2 + 1'301^2).(2'954^2 + 1'788^2 + 3'903^2)}} =$$

$$= \frac{9'441}{1'968.5'208} = 0'921$$

$$SIM(D2,Q) = \frac{3'556.1'778 + 10'408.3'903}{\sqrt{(3'556^2 + 10'408^2).5'208}} = \frac{46'945}{10'999.5'208} = 0'820$$

$$SIM(D3,Q) = \frac{7'386.2'954 + 1'778.1'778}{\sqrt{(7'386^2 + 1'778^2).5'208}} = \frac{24'980}{7'597.5'208} = 0'631$$

En consecuencia, el sistema ordenaría los documentos por orden decreciente de similitud con la consulta; en nuestro caso, pues, la respuesta del SRI sería:

D1

D2

D3

B) Si el SRI emplease el coeficiente de Jaccard para el cálculo de similitudes, tendríamos:

$$SIM(D1,Q) = \frac{\sum_{k=1}^{t}(TERMik.TERMjk)}{\sum_{k=1}^{t}TERMik^2 + \sum_{k=1}^{t}TERMjk^2 - \sum_{k=1}^{t}(TERMik.TERMjk)} =$$

$$\frac{(1'477.2'954 + 1'301.3'903)}{(1'477^2 + 1'301^2) + (2'954^2 + 1'778^2 + 3'903^2) - (1'477.2'954 + 1'301.3'903)} =$$

$$\frac{9'441}{3'874 + 27'121 - 9'441} = \frac{9'441}{21'554} = 0'438$$

$$SIM(D2,Q) = \frac{3'556.1'778 + 10'408.3'903}{(3'556^2 + 10'408^2) + (2'954^2 + 1'778^2 + 3'903^2) - 46'945} =$$

$$\frac{46'945}{120'972 + 27'121 - 46'945} = \frac{46'945}{101'148} = 0'464$$

$$SIM(D3,Q) = \frac{7'386.2'954 + 1'778.1'778}{(7'386^2 + 1'778^2) + (2'954^2 + 1'778^2 + 3'903^2) - 24'980} =$$

$$\frac{24'980}{57'714 + 27'121 - 24'980} = \frac{24'980}{59'855} = 0'417$$

Ordenando de nuevo por orden decreciente de similitud obtendríamos la respuesta del sistema:

D2

D1

D3

Como podemos observar, la respuesta es distinta. El coeficiente de Jaccard tiende a disminuir las diferencias de similitud (una diferencia máxima de 0’047 entre los documentos D1, D2 y D3), mientras que el coeficiente del coseno tiende a aumentar dichas diferencias (una diferencia máxima de 0’29 entre los documentos D1, D2 y D3). Este efecto puede llegar a provocar, como hemos comprobado en nuestro caso, una modificación en la posición relativa de los documentos de la colección y, en consecuencia, ligeras diferencias en la respuesta del sistema ante una consulta dependiendo del coeficiente de similitud empleado.

Ejercicio 3

Sea un SRI cuyo fichero diccionario es el siguiente:

Fichero diccionario

Término	Nº total docs.	Frec. Abs.	Lista
t_1	2	3	1/2, 2/1
t_2	1	5	4/5
t_3	4	12	1/1, 2/5, 3/3, 4/3
t_4	3	7	1/5, 2/1, 4/1
t_5	2	2	2/1, 3/1

A) Halle la respuesta del sistema (conforme al modelo vectorial básico, empleando ponderación tf.idf y coeficiente del coseno) ante la consulta:

$$Q = < 5,\ 0,\ 3,\ 0,\ 1 >$$

B) Represente los documentos de la colección y halle la respuesta del sistema (conforme al modelo booleano básico) ante la consulta:

$$Q = t_1 \text{ AND } t_3 \text{ AND } t_5$$

C) Halle la respuesta del sistema (conforme al modelo booleano básico) ante la consulta:

$$Q = t_1 \text{ OR } t_3 \text{ OR } t_5$$

D) Represente los documentos de la colección y halle la respuesta del sistema (conforme al modelo probabilístico básico) ante la consulta:

Q= < 1, 0, 1, 0, 1 >

Solución

(A)

Para hallar la respuesta del sistema conforme al modelo vectorial, obtendremos en primer lugar la descripción de los documentos de la colección conforme al factor tf:

D1 = < 2, 0, 1, 5, 0 >

D2 = < 1, 0, 5, 1, 1 >

D3 = < 0, 0, 3, 0, 1 >

D4 = < 0, 5, 3, 1, 0 >

A continuación, calculamos los valores del factor IDF para cada uno de los términos del sistema:

$$IDF1 = \log(\frac{N}{DOCFREQ1}) + 1 = \log(\frac{4}{2}) + 1 = \log 2 + 1 = 1'301$$

$$IDF2 = \log(\frac{N}{DOCFREQ2}) + 1 = \log(\frac{4}{1}) + 1 = 1'602$$

$$IDF3 = \log(\frac{N}{DOCFREQ3}) + 1 = \log(\frac{4}{4}) + 1 = \log 1 + 1 = 1$$

$$IDF4 = \log(\frac{N}{DOCFREQ4}) + 1 = \log(\frac{4}{3}) + 1 = 1'125$$

$$IDF5 = \log(\frac{N}{DOCFREQ5}) + 1 = \log(\frac{4}{2}) + 1 = \log 2 + 1 = 1'301$$

Luego las representaciones de los cuatro documentos del sistema (empleando ponderación tf.idf) serían:

D1 = < 2'602, 0 , 1, 5'625, 0 >

D2 = < 1'301, 0 , 5, 1'125, 1'301 >

D3 = < 0 , 0 , 3, 0 , 1'301 >

D4 = < 0 , 8'01, 3, 1'125, 0 >

Una vez hallada la representación de los documentos de la colección podemos hallar la similitud de cada uno de ellos con la consulta Q=<5,0,3,0,1> mediante el coeficiente del coseno:

$$SIM(D1,Q) = \frac{\sum_{k=1}^{t}(TERMik.TERMjk)}{\sqrt{\sum_{k=1}^{t}(TERMik)^2.\sum_{k=1}^{t}(TERMjk)^2}} = \frac{2'602.5 + 3}{\sqrt{(2'602^2 + 1 + 5'625^2).\sqrt{(5^2 + 3^2 + 1^2)}}} =$$

$$= \frac{16'010}{6'278.5'916} = 0'431$$

$$SIM(D2,Q) = \frac{5.1'301 + 5.3 + 1'301}{\sqrt{(1'301^2 + 25 + 1'125^2 + 1'301^2).5'916}} = \frac{22'806}{5'445.5'916} = 0'708$$

$$SIM(D3,Q) = \frac{3.3 + 1.1'301}{\sqrt{(3^2 + 1'301^2).5'916}} = \frac{10'301}{3'270.5'916} = 0'532$$

$$SIM(D4,Q) = \frac{3.3}{\sqrt{(8'01^2 + 9 + 1'125^2).5'916}} = \frac{9}{8'627.5'916} = 0'176$$

En consecuencia, la respuesta del sistema sería:

D2

D3

D1

D4

(B)

Conforme al modelo booleano, la representación de los documentos de la colección sería la siguiente:

D1 = < 1, 0, 1, 1, 0 >

D2 = < 1, 0, 1, 1, 1 >

D3 = < 0, 0, 1, 0, 1 >

D4 = < 0, 1, 1, 1, 0 >

Para hallar la respuesta del sistema ante la consulta

Q= t_1 AND t_3 AND t_5

hallamos primeramente los conjuntos correspondientes a los términos presentes en la consulta:

t_1= {D1,D2}

t_3= {D1,D2,D3,D4}

t_5= {D2,D3}

Luego la respuesta sería:

R = {D1,D2} ∩ {D1,D2,D3,D4} ∩ {D2,D3} = {D1,D2} ∩ {D2,D3}= D2

Es decir, la respuesta constaría únicamente del documento D2.

(C)

La respuesta del sistema, si la consulta fuese Q= t_1 OR t_3 OR t_5, sería:

R = {D1,D2} U {D1,D2,D3,D4} U {D2,D3} = {D1,D2,D3,D4} U {D2,D3}=

R = {D1,D2,D3,D4}

Esto es, la colección en su integridad, sin imponer un orden entre ellos:

D1, D2, D3, D4 [en cualquier orden]

(D)

Si el SRI se basa en el modelo probabilístico, la descripción de los documentos es también binaria, como en el modelo booleano. Por tanto:

D1 = < 1, 0, 1, 1, 0 >

D2 = < 1, 0, 1, 1, 1 >

D3 = < 0, 0, 1, 0, 1 >

D4 = < 0, 1, 1, 1, 0 >

Para hallar la respuesta inicial del sistema a la consulta

Q= < 1, 0, 1, 0, 1 >

Partimos de la siguiente hipótesis inicial:

$p(t_1/R) = 0'5$ $\quad p(t1/\overline{R}) = \frac{2}{4} = 0'5$

$p(t_3/R) = 0'5$ $\quad p(t3/\overline{R}) = \frac{4}{4} = 1$

$p(t_5/R) = 0'5$ $\quad p(t5/\overline{R}) = \frac{2}{4} = 0'5$

Por otra parte, los cálculos de la similitud entre los documentos de la colección y la consulta Q son los siguientes:

Q= < 1, 0, 1, 0, 1 >

D1 = < 1, 0, 1, 1, 0 >

D2 = < 1, 0, 1, 1, 1 >

D3 = < 0, 0, 1, 0, 1 >

D4 = < 0, 1, 1, 1, 0 >

sim(D1,Q) = $c(t_1) + c(t_3)$

sim(D2,Q) = $c(t_1) + c(t_3) + c(t_5)$

sim(D3,Q) = $c(t_3) + c(t_5)$

sim(D4,Q) = $c(t_3)$

A su vez, los valores de los coeficientes correspondientes a los términos presentes en la consulta son:

$$c(t1) = \log(\frac{0'5}{1-0'5}) + \log(\frac{1-0'5}{0'5}) = 0 + 0 = 0$$

$$c(t3) = \log(\frac{0'5}{1-0'5}) + \log(\frac{1-1}{1}) = 0 + \log 0 = -\infty$$

$$c(t5) = \log(\frac{0'5}{1-0'5}) + \log(\frac{1-0'5}{0'5}) = 0 + 0 = 0$$

En efecto, conforme a la lógica del modelo probabilístico, el término t_3 aparece en todos los documentos irrelevantes (esa es la hipótesis inicial en nuestro caso), de manera que el coeficiente correspondiente a dicho término sería el más negativo posible desde el punto de vista matemático, esto es, (menos infinito).

Si mantenemos el valor del coeficiente del término t_3, todas las similitudes valdrían , imposibilitando la ordenación de los documentos de la colección. Dado que dicho

componente es común a todas las similitudes, podemos eliminar el componente correspondiente al coeficiente del término t_3 en todas las similitudes de todos los documentos de la colección. De esta forma:

$sim(D1,Q) = c(t_1)$

$sim(D2,Q) = c(t_1) + c(t_5)$

$sim(D3,Q) = c(t_5)$

$sim(D4,Q) = 0$

Sin embargo, en este caso tanto el coeficiente correspondiente a t_1 como el correspondiente a t_5 valen cero, de manera que todas las similitudes de todos los documentos de la colección siguen presentando el mismo valor ( antes, 0 ahora):

$sim(D1,Q) = 0$

$sim(D2,Q) = 0$

$sim(D3,Q) = 0$

$sim(D4,Q) = 0$

En definitiva, la respuesta del sistema sería:

D1, D2, D3, D4 [en cualquier orden]

Como vemos, en este caso tan particular (extremadamente extraño en la práctica) el modelo probabilístico daría una respuesta semejante al modelo booleano, sin poder ordenar los documentos de la colección. Tan solo el modelo vectorial sería capaz de imponer un orden en la respuesta.

Ejercicio 4

Sea un sistema compuesto por 100 documentos caracterizados por 4 términos de indización con el siguiente fichero diccionario (en relación a los tres primeros documentos):

Término	Nº docs	Frec. Abs.	Lista
t_1	10	20	1/2, 3/2
t_2	20	38	2/3, 3/1
t_3	25	26	1/1, 2/1
t_4	5	15	2/2, 3/4

A) Conforme al modelo probabilístico, calcule la respuesta inicial del sistema en relación a los tres primeros documentos y la consulta

Q= <1,0,1,0>

B) Posteriormente se pide al usuario que evalúe los resultados. Señala 10 documentos como relevantes, con la siguiente distribución:

t_1 aparece en 2 documentos relevantes

t_2 aparece en los 10 documentos relevantes

t_3 aparece en 8 documentos relevantes

t_4 aparece en 3 documentos relevantes

Hallar la nueva respuesta del sistema.

C) Conforme al modelo vectorial (empleando ponderación tf.idf), calcule la respuesta del sistema en relación a D1, D2 y D3 mediante el coeficiente del coseno ante la misma consulta Q=<1,0,1,0>.

D) Conforme al modelo booleano, hallar la respuesta del sistema a la consulta equivalente:

$$Q = t_1 \text{ OR } t_3$$

Solución

A) El SRI efectúa una hipótesis inicial sobre la distribución de los términos (únicamente aquellos presentes en la consulta) en los conjuntos de documentos relevantes e irrelevantes de la colección respecto a la consulta Q:

$$p(t1/R) = 0'5 \qquad p(t1/\bar{R}) = \frac{n1}{N} = \frac{10}{100} = 0'1$$

$$p(t3/R) = 0'5 \qquad p(t3/\bar{R}) = \frac{n3}{N} = \frac{25}{100} = 0'25$$

A continuación el SRI efectúa el cálculo de los coeficientes o pesos iniciales correspondientes a los términos presentes en la consulta:

$$c1 = \log\frac{0'5}{0'5} + \log\frac{1-0'1}{0'1} = \log 9$$

$$c3 = \log\frac{1-0'25}{0'25} = \log\frac{\frac{3}{4}}{\frac{1}{4}} = \log 3$$

Luego el SRI calcula la similitud entre cada documento de la colección y la consulta Q mediante la suma de los coeficientes o pesos correspondientes a los términos presentes simultáneamente en el documento y en la consulta:

$$sim(D1,Q) = c1 + c3 = \log 9 + \log 3$$
$$sim(D2,Q) = c3 = \log 3$$
$$sim(D3,Q) = c1 = \log 9$$

En consecuencia, el SRI mostraría, en su respuesta, los documentos ordenados en orden decreciente de su probabilidad de relevancia en relación a la consulta:

D1

D3

D2

B) El SRI calcula de nuevo la distribución de los términos de la consulta en los conjuntos de relevantes e irrelevantes definidos por el usuario:

$$p(t1/R) = \frac{|V1|}{|V|} = \frac{2}{10} \qquad p(t1/\bar{R}) = \frac{n1 - |V1|}{N - |V|} = \frac{10-2}{100-10} = \frac{8}{90}$$

$$p(t3/R) = \frac{|V3|}{|V|} = \frac{8}{10} \qquad p(t3/\bar{R}) = \frac{n3 - |V3|}{N - |V|} = \frac{25-8}{90} = \frac{17}{90}$$

El SRI calcula los nuevos coeficientes o pesos correspondientes a los términos presentes en la consulta, en base al juicio de relevancia emitido por el usuario:

$$c1 = \log \frac{\frac{2}{10}}{1-\frac{2}{10}} + \log \frac{\frac{90-8}{90}}{\frac{8}{90}} = \log \frac{1}{4} + \log \frac{82}{8} = -0'602 + 1'011 = 0'409$$

$$c3 = \log \frac{\frac{8}{10}}{1-\frac{8}{10}} + \log \frac{1-\frac{17}{90}}{\frac{17}{90}} = \log 4 + \log \frac{73}{17} = 0'602 + 0'633 = 1'235$$

El SRI calcula de nuevo la similitud entre cada documento de la colección y la consulta Q, mediante la suma de los coeficientes o pesos correspondientes a los términos presentes simultáneamente en el documento y en la consulta:

$$sim(D1,Q) = c1 + c3 = 0'409 + 1'235$$
$$sim(D2,Q) = c3 = 1'235$$
$$sim(D3,Q) = c1 = 0'409$$

Finalmente, el SRI mostraría en su respuesta los documentos en orden decreciente de probabilidad de relevancia en relación a la consulta:

D1

D2

D3

Como vemos, D2 ha subido en posición relativa. Ello es debido a que D2 contiene el término t_3, cuya presencia en los documentos relevantes (juzgados así por el usuario) es muy elevada (8 de 10); en cambio, D3 contiene el término t_1, cuya presencia en los documentos relevantes (juzgados así por el usuario) es minoritaria (2 de 10).

C) Para hallar la respuesta del sistema conforme al modelo vectorial, obtendremos en primer lugar la descripción de aquellos documentos de la colección de los que conocemos sus datos, conforme al factor tf:

D1 = < 2, 0, 1, 0 >

D2 = < 0, 3, 1, 2 >

D3 = < 2, 1, 0, 4 >

A continuación, calculamos los valores del factor IDF para cada uno de los términos del sistema:

$$IDF1 = \log(\frac{N}{DOCFREQ1}) + 1 = \log(\frac{100}{10}) + 1 = \log 10 + 1 = 2$$

$$IDF2 = \log(\frac{N}{DOCFREQ2}) + 1 = \log(\frac{100}{20}) + 1 = 1'699$$

$$IDF3 = \log(\frac{N}{DOCFREQ3}) + 1 = \log(\frac{100}{25}) + 1 = \log 4 + 1 = 1'602$$

$$IDF4 = \log(\frac{N}{DOCFREQ4}) + 1 = \log(\frac{100}{5}) + 1 = 2'301$$

Luego las representaciones de los tres documentos del sistema de los que conocemos sus datos (empleando ponderación tf.idf) serían:

D1 = < 4 , 0 , 1’602 , 0 >

D2 = < 0 , 5’097 , 1’602 , 4’602 >

D3 = < 4 , 1’699 , 0 , 9’204 >

Una vez hallada la representación de los documentos podemos hallar la similitud de cada uno de ellos con la consulta Q=<1,0,1,0> mediante el coeficiente del coseno:

$$SIM(D1,Q) = \frac{\sum_{k=1}^{t}(TERMik.TERMjk)}{\sqrt{\sum_{k=1}^{t}(TERMik)^2.\sum_{k=1}^{t}(TERMjk)^2}} = \frac{4+1'602}{\sqrt{(4^2+1'602^2).(1^2+1^2)}} =$$

$$= \frac{5'602}{\sqrt{18'566}.\sqrt{2}} = \frac{5'602}{6'094} = 0'919$$

$$SIM(D2,Q) = \frac{1'602}{\sqrt{(5'097^2+1'602^2+4'602^2).2}} = \frac{1'602}{9'972} = 0'161$$

$$SIM(D3,Q) = \frac{4}{\sqrt{(4^2+1'699^2+9'204^2).2}} = \frac{4}{14'394} = 0'278$$

En consecuencia, la respuesta del sistema sería:

D1

D3

D2

Como podemos observar, la misma respuesta que la inicial del modelo probabilístico.

D) Conforme al modelo booleano, para hallar la respuesta del sistema ante la consulta Q= t_1 OR t_3, hallamos primeramente los conjuntos correspondientes a los términos presentes en la consulta:

t_1= {D1,D3}

t_3= {D1,D2}

Luego la respuesta sería:

R = {D1,D3} U {D1,D2} = {D1,D2,D3}

Esto es, todos los documentos de la colección de los que tenemos datos, sin imponer un orden entre ellos:

D1, D2, D3 [en cualquier orden]