Modelos Clásicos de RI

UCR – ECCI

CI-2414 Recuperación de Información

Prof. Kryscia Daviana Ramírez Benavides

Características de los Modelos Clásicos

- Los documentos se describen a través de un conjunto de términos representativos llamados **términos índice**.
- Los índices son principalmente sustantivos y adjetivos, y se usan en menor medida verbos, adjetivos, advervios, etc.
- Sin embargo, se pueden considerar todos los términos como importantes en una aproximación llamada *full text*.
- No todos los términos son igualmente importantes. Por ejemplo: un término que aparece en todos los documentos de una colección será menos importante que otro que aparezca sólo en unos pocos, puesto que ayuda a discernir.

Características de los Modelos Clásicos (cont.)

- El proceso de decidir la importancia de un término se puede realizar a través de la asignación de **pesos**.
  - Para k_i (término), d_j (documento), $w_{ij} \geq 0$ es el peso asociado a l término en el documento.
- Los pesos de los términos son mutuamente independientes, esto es, sabiendo el peso w_{ij}, no podemos saber nada a priori del peso w_{i+1j}.

Definición Formal de los Modelos Clásicos

- Sea t el número de términos índice en el sistema, y k_i un término índice genérico. $K = \{k_1,...,k_t\}$ es el conjunto de índices. Un peso $w_{ij} > 0$ se asocia con cada término k_i del documento d_j. Para un término que no aparece en el documento, $w_{ij} = 0$. Con cada documento d_j hay asociado un vector de índices $d_j = (w_{1j}, w_{2j},..., w_{tj})$. Además, definimos una función g_i que devuelve el peso asociado con índice k_i en un vector t-dimensional: $g_i(d_j) = w_{ij}$.

Nomenclatura General

- q → consulta
- d_j → documento
- k_i → término del índice genérico
- w_{ij} → peso de relevancia; i = término, j = documento
- $sim(d_j,q)$ → función de similitud del documento j con la consulta q
- $g_i(d_j)$ → función que retorna el peso asociado con el término índice k_i en un vector t-dimensional ($g_i(d_j) = w_{ij}$)

Modelo Booleano

Características del Modelo Booleano

- Modelo clásico basado en la teoría de conjuntos y el álgebra de Boole.
- Es el modelo más simple, por lo cuál es adoptado por muchos SRI tempranos.
- La relevancia es binaria: un documento es relevante o no lo es.
- Consultas de una palabra: un documento relevante si contiene la palabra.
- Diseñado para recuperar todos los registros almacenados, que contengan la combinación exacta de palabras claves incluidas en la consulta.

Características del Modelo Booleano (cont.)

- Los documentos se representan por conjuntos de términos contenidos en ellos.
- Las consultas se expresan como expresiones booleanas con una semántica clara y concreta. Se busca una representación óptima a través de una FND (Forma Normal Disjunta).
- Consultas AND: los documentos deben contener todas las palabras.
- Consultas OR: los documentos deben contener alguna palabra.
- Consultas NOT: los documentos no deben contener la palabra.

Características del Modelo Booleano (cont.)

- Modelo más primitivo y bastante malo para RI.
- Es bastante popular.
- Los documentos se encuentran representados por conjuntos de palabras clave, generalmente almacenadas en un fichero inverso.
- Se basa más en recuperación de datos que en recuperación de información.

Caracterización Formal del Modelo Booleano

$$[\ D, Q, F, R(q_i, d_j)\ ]$$

- **D** = Conjunto de vectores que representan a los documentos, formado por los pesos (0 ó 1) de los términos de la colección en el documento.
- **Q** = Conjunto de vectores que representan a las consultas, formado por los pesos (0 ó 1) de los términos de la colección en la consulta.
- ***F*** = Álgebra Booleana.
- $\boldsymbol{R(q_i, d_j)}$ = Forma Normal Disjunta (FND, q_{cc} y q_{fnd}).

Nomenclatura

- a,b,c $\rightarrow$ términos índice o palabras clave
- q_{fnd} $\rightarrow$ forma normal disjunta de la consulta
- q_{cc} $\rightarrow$ componente conjutivo de la consulta

Definición Formal del Modelo Booleano

- Los pesos de los términos son binarios ($w_{ij} \in \{0,1\}$). Una consulta es una expresión booleana convencional. Si q_{fnd} es la forma normal disjunta de una consulta, y q_{cc} alguno de los componentes de esta **fnd**, la similitud de un documento d_j con una consulta q se define como:

$$sim(d_j,q) = \begin{cases} \mathbf{1} \text{ si } \exists\, q_{cc} \mid (q_{cc} \in q_{fnd}) \wedge (\forall k_i,\, g_i(d_j) = g_i(q_{cc})) \\ \mathbf{0} \text{ en otro caso} \end{cases}$$

- Si $sim(d_j,q)=1$, entonces el documento se predice como relevante. En cualquier otro caso, el documento no es relevante.

Ejemplo del Modelo Booleano

- Consulta Genérica:

$$q = k_a \wedge (k_b \vee \neg k_c)$$

- Consulta en FND:

$$q = (k_a \wedge k_b) \vee (k_a \wedge \neg\, k_c)$$

$$q_{fnd} = (k_a \wedge k_b \wedge k_c) \vee (k_a \wedge k_b \wedge \neg k_c) \vee (k_a \wedge \neg k_b \wedge \neg k_c)$$

$$q_{fnd} = (1,1,1) \vee (1,1,0) \vee (1,0,0)$$

q_{cc}

Ejemplo del Modelo Booleano (cont.)

$d_1 = (0,0,1)$

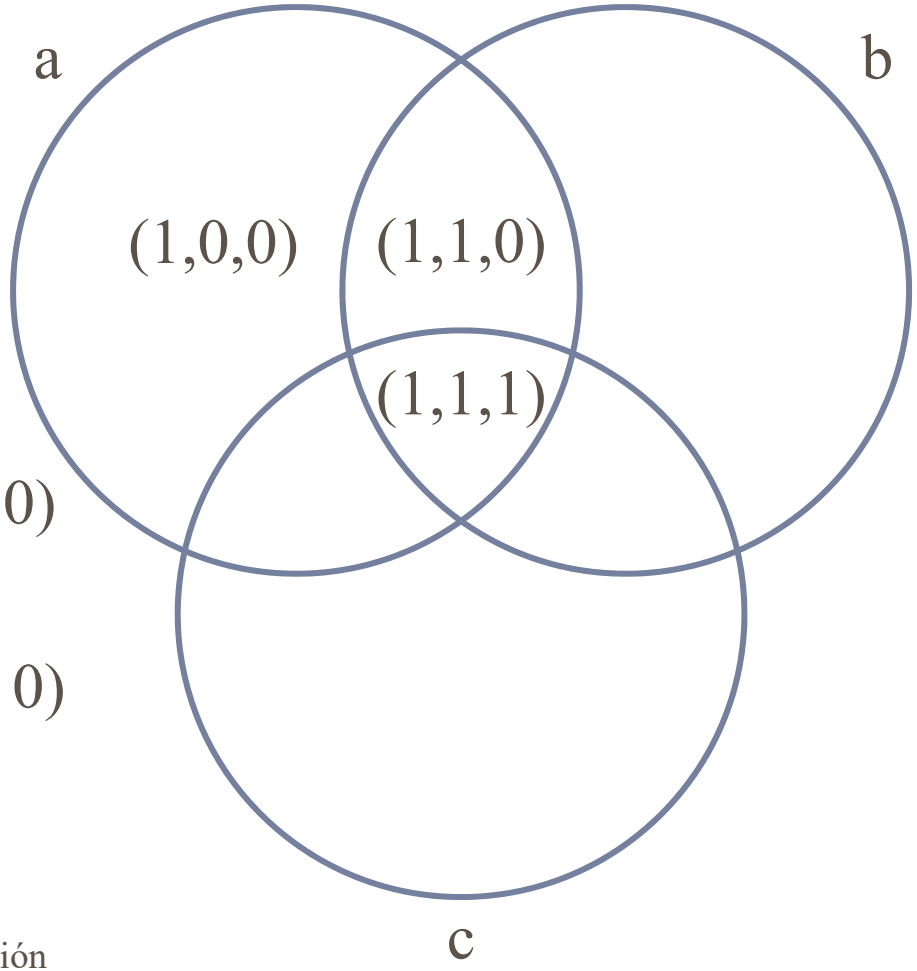
$d_2 = (0,1,0)$

$d_3 = (0,1,1)$

$d_4 = (1,0,1)$

$q = k_a \wedge (k_b \vee \neg k_c)$

$q_{fnd} = (1,1,1) \vee (1,1,0) \vee (1,0,0)$



No hay respuesta parcial (1 ó 0)

Resultado: $sim(d_j,q) = 0$

Ventajas y Razones de Popularidad del Modelo Booleano

- Más sencillo imposible.
- Es simple de formalizar y eficiente de implementar.
- En situaciones operacionales alcanza un alto estándar de desempeño.
- Es de las primeras ideas que a uno se le ocurren.
- Muchos de los primeros SRI se basaron en él.
- En algunos casos (usuarios expertos) puede ser adecuado.
- Puede ser útil en combinación con otro modelo, por ejemplo: excluir documentos.
- Puede ser útil con mejores interfaces.

Desventajas del Modelo Booleano

- No discrimina entre documentos más y menos relevantes.
- Da lo mismo que un documento contenga una o mil veces las palabras de la consulta.
- Todos los términos incluidos en la pregunta o los documentos tienen igual importancia.
- Da lo mismo que cumpla *una o todas las cláusulas de un OR*.
- No considera un calce parcial de un documento, por ejemplo: que cumpla con *casi todas las cláusulas de un AND*.

Desventajas del Modelo Booleano (cont.)

- No permite siquiera ordenar los resultados. Los resultados obtenidos nos están clasificados en ningún orden de importancia para el usuario.
- El tamaño de la respuesta obtenida en respuesta a una consulta es difícil de controlar.
- Aunque las expresiones booleanas tienen una semántica precisa, no es sencillo trasladar las necesidades de información de un usuario a expresiones booleanas.

Desventajas del Modelo Booleano (cont.)

- El usuario promedio no lo entiende, por ejemplo, ante la necesidad de información:
  - "Necesito investigar sobre los Aztecas y sobre los Incas"

    se convierte en
  - Aztecas AND Incas

  grave error, se perderán excelentes documentos que traten una sola de las culturas en profundidad, debió ser:
  - Aztecas OR Incas

Modelo Vectorial

Características del Modelo Vectorial

- Se selecciona un conjunto de palabras útiles para discriminar (términos índice o *keywords*).
- En los sistemas modernos, toda palabra del texto es un término, excepto posiblemente las palabras vacías o *stopwords* (artículos, conjunciones, preposiciones).
- Se puede enriquecer con procesos de **lematización** (*stemming*), **etiquetado** e **identificación de frases**.
- Asume que el uso de pesos binarios es limitativo y propone un marco con posibilidad de relevancia parcial, para recuperar documentos con una coincidencia parcial.

Características del Modelo Vectorial (cont.)

- Por tanto, se asignan pesos no binarios a los términos en los documentos.
- En lugar de predecir si un documento es o no relevante, se proporciona un grado de relevancia.
- Se pretende computar el grado de similitud entre documentos y consultas de forma gradual, y no absoluta.
- Se establece un umbral de relevancia para decidir cuando mostrar un documento como relevante.
- El problema para obtener la relevancia consistirá en la forma de asignar pesos.

Características del Modelo Vectorial (cont.)

- El resultado será un conjunto de documentos respuesta a una consulta ordenados por un ranking de relevancia.
- En particular, una consulta se puede ver como un documento (formado por esas palabras) y por lo tanto como un vector.
- En este modelo se tiene un conjunto de términos (k_1, k_2, ..., k_t) y un conjunto de documentos (d_1, d_2, ..., d_N).
- El modelo es más general y permite cosas como:
  - La consulta sea un documento.
  - Hacer *clustering* de documentos similares.
  - Retroalimentación por relevancia (*relevance feedback:* "*more like this*").
- El modelo más popular en RI hoy en día, ya que es el más utilizado.

Caracterización Formal del Modelo Vectorial

$$[\ D, Q, F, R(q_i,d_j)\ ]$$

- **D** = Espacio vectorial de documentos, formado por los pesos de los términos de la colección en el documento.
- **Q** = Espacio vectorial de consultas, formado por los pesos de los términos de la colección en la consulta.
- ***F*** = Álgebra Vectorial.
- $\boldsymbol{R(q_i,d_j)}$ = Distancia Coseno (Coseno del Ángulo entre los Vectores)

Nomenclatura

- $freq_{ij}$ $\rightarrow$ frecuencia del término en el documento
- $max_l freq_{lj}$ $\rightarrow$ máxima frecuencia de un término en el documento
- f_{ij} $\rightarrow$ frecuencia normalizada del término en el documento
- idf_i $\rightarrow$ frecuencia inversa del término dentro de los documentos de la colección
- w_{iq} $\rightarrow$ peso de relevancia;
  i = término, q = consulta

Nomenclatura (cont.)

- N $\rightarrow$ total de documentos de una colección
- n_i $\rightarrow$ los documentos en los que aparece el término en el documento

Definición Formal del Modelo Vectorial

- El peso w_{ij} que se asocia a un par (k_i, d_j) es positivo y no binario. De igual modo, los pesos de los términos en una consulta se someten a los mismos pesos, de modo que $w_{iq} \geq 0$ es el peso asociado asociado al par (k_i,q). El vector q se define como $q = (w_{1q}, w_{2q}, \ldots, w_{tq})$ siendo t el número total de términos indexados en el sistema. De igual forma, el vector documento se representa por $\rightarrow d_j = (w_{1j}, w_{2j}, \ldots, w_{tj})$.
- Por tanto, un documento y una consulta se representan como vectores t-dimensionales (vectores en un espacio de t dimensiones, siendo t el número de términos indexados en la colección de documentos).

Definición Formal del Modelo Vectorial (cont.)

- Un documento d_j se modela como un vector

$$\rightarrow$$

$$d_j \rightarrow d_j = (w(k_1, d_j), \ldots, w(k_t, d_j))$$

donde $w(k_i, d_j)$ es el peso del término k_i en el documento d_j.

- Una consulta q se modela como un vector

$$\rightarrow$$

$$q \rightarrow q = (w(k_1, q), \ldots, w(k_t, q))$$

donde $w(k_i, q)$ es el peso del término k_i en el documento q.

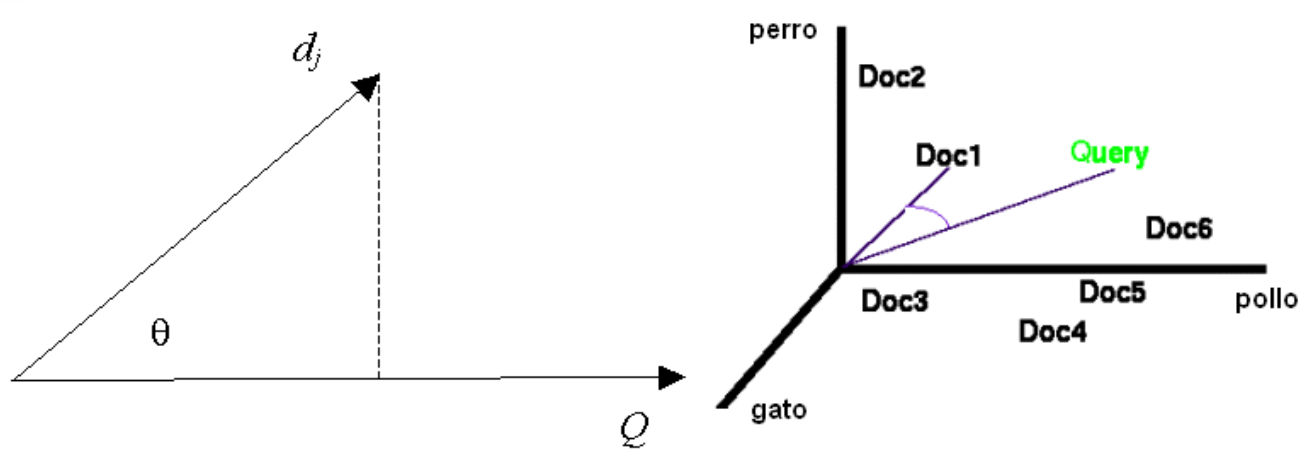
Definición Formal del Modelo Vectorial (cont.)

- La similaridad entre el documento d_j y la consulta q se toma como la correlación entre sus vectores, y puede ser cuantificada por el coseno del ángulo entre ellos (la función coseno normaliza los vectores respecto a su longitud):

$$sim(d_j, q) = \frac{\vec{d}_j \bullet \vec{q}}{\left\|\vec{d}_j\right\| \times \left\|\vec{q}\right\|} = \frac{\sum_{i=1}^{t} \left(w_{ij} \times w_{iq}\right)}{\sqrt{\sum_{i=1}^{t} (w_{ij})^2} \times \sqrt{\sum_{i=1}^{t} (w_{iq})^2}}$$

Definición Formal del Modelo Vectorial (cont.)

- La similaridad es un número entre 0 y 1, pues así son los pesos de los términos de los vectores:
  - Si $sim(d_j,q) = 1$, entonces el documento y la consulta son iguales (vectores paralelos).
  - Si $sim(d_j,q) = 0$, entonces el documento y la consulta no comparten términos (vectores ortogonales).



Definición Formal del Modelo Vectorial (cont.)

- Sobre la fórmula del coseno:
  - La norma del vector consulta no afecta al ranking porque es igual para todos los documentos, cosa que no pasa con la norma del vector documento.
- Problema de *clustering* en RI: definir que documentos son relevantes y que documentos no lo son. Se pueden usar dos medidas para ello:
  - *Similitud intra-cluster:* Se puede utilizar como medida la frecuencia de términos (*tf*).
  - *Diferencia inter-cluster:* Se puede utilizar como medida la frecuencia de documento inversa (*idf*).
- Estas medidas (*tf* y *idf*) se utilizan para el cálculo de los pesos de los términos.

Definición Formal del Modelo Vectorial (cont.)

- Los pesos de los términos pueden ser calculados de distintos modos, aunque el modo más común es el conocido por *tf-idf*:
  - Factor *tf* $\Rightarrow$ Es la frecuencia del término k_i dentro del documento d_j. Mide la calidad del término como descriptor del documento.
  - Factor *idf* $\Rightarrow$ Frecuencia inversa del término k_i dentro de los documentos de la colección. Aquellos términos que aparecen en muchos documentos no son útiles para distinguir entre documentos relevantes y no relevantes.
- Se evalúa lo importante que es el término en el documento por lo importante que es el término en la colección de documentos.

Definición Formal del Modelo Vectorial (cont.)

- Si un término aparece mucho en un documento, se supone que es importante en ese documento (*tf* crece).
- Pero si aparece en muchos documentos, entonces no es útil para distinguir ningún documento de los otros (*idf* decrece).
- Se normalizan los módulos de los vectores para no favorecer documentos más largos.
- Lo que se intenta medir es cuánto ayuda ese término a distinguir ese documento de los demás.

Definición Formal del Modelo Vectorial (cont.)

- Sea N el total de documentos de una colección, y n_i los documentos en los que aparece el término k_i.
- La frecuencia del término k_i en el documento d_j la denotamos por $freq_{ij}$.
- La frecuencia normalizada del término k_i en el documento d_j es f_{ij}.
- El máximo se obtiene sobre los términos del documento.
- La frecuencia de documento inversa será idf_i.

Definición Formal del Modelo Vectorial (cont.)

- La frecuencia normalizada f_{ij} del término k_i en el documento d_j se calcula como

$$f_{ij} = \frac{freq_{ij}}{\max_l \ freq_{lj}}$$

donde $freq_{ij}$ es la frecuencia del término k_i en el documento d_j, y el máximo se calcula sobre todos los términos que aparecen en el documento d_j (frecuencia del término más frecuente en el documento d_j).

Definición Formal del Modelo Vectorial (cont.)

- La frecuencia inversa del término k_i se calcula como

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

donde N es el número total de documentos en la colección y n_i el número de documentos donde aparece el término k_i.

Definición Formal del Modelo Vectorial (cont.)

- Los pesos de los términos serán calculados como

$$w_{ij} = f_{ij} \times idf_i = \frac{freq_{ij}}{\max_l freq_{lj}} \times \log \frac{N}{n_i}$$

o variaciones de la misma.

- Para los pesos de la consulta se sugiere el siguiente cálculo:

$$w_{iq} = \left(0.5 + 0.5 \times f_{iq}\right) \times idf_i = \left(0.5 + 0.5 \times \frac{freq_{iq}}{\max_l freq_{lq}}\right) \times \log \frac{N}{n_i}$$

Ejemplo del Modelo Vectorial

- Se tiene siete artículos sobre el tema de Comida, en donde se distinguen por usar solamente tres palabras útiles para discriminar: postres, panes y vegetales. Los artículos hablan sobre:
  - Primer artículo (d1) → postres.
  - Segundo artículo (d2) → panes.
  - Tercer artículo (d3) → panes y vegetales.
  - Cuarto artículo (d4) → postres, panes y vegetales.
  - Quinto artículo (d5) → postres y panes, más de postres.
  - Sexto artículo (d6) → postres y panes de igual forma.
  - Sétimo artículo (d7) → postres y panes, más de panes.

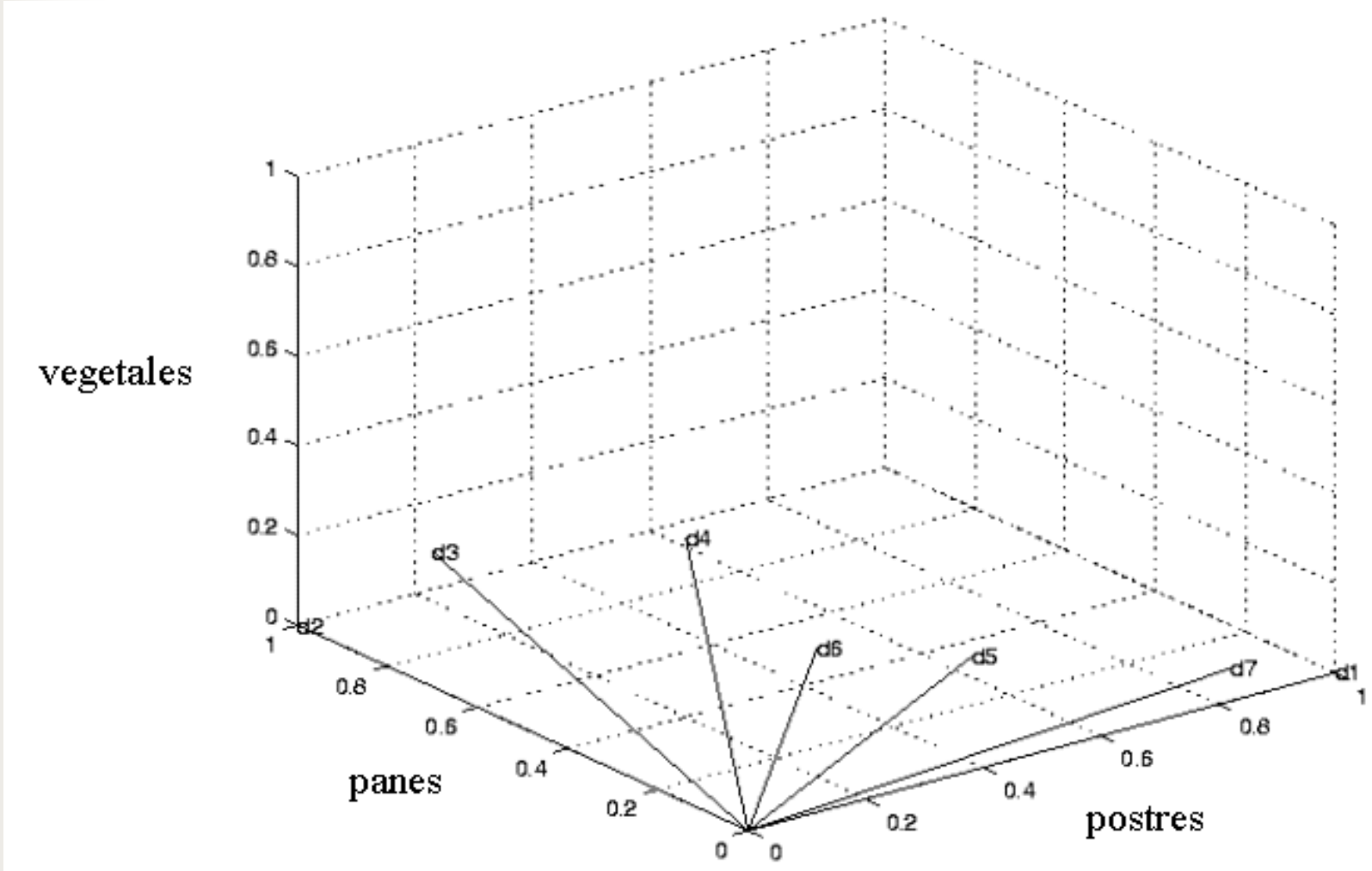
Ejemplo del Modelo Vectorial (cont.)

- El vector que se utilizará tanto para documentos y consultas se define como (postres, panes, vegetales). Los vectores documentos que se generan son:
  - Primer artículo (d1) → (0.146,0,0).
  - Segundo artículo (d2) → (0,0.067,0).
  - Tercer artículo (d3) → (0,0.067,0.544).
  - Cuarto artículo (d4) → (0.146,0.67,0.544).
  - Quinto artículo (d5) → (0.146,0.033,0).
  - Sexto artículo (d6) → (0.146,0.067,0).
  - Sétimo artículo (d7) → (0.029,0.067,0).

Ejemplo del Modelo Vectorial (cont.)

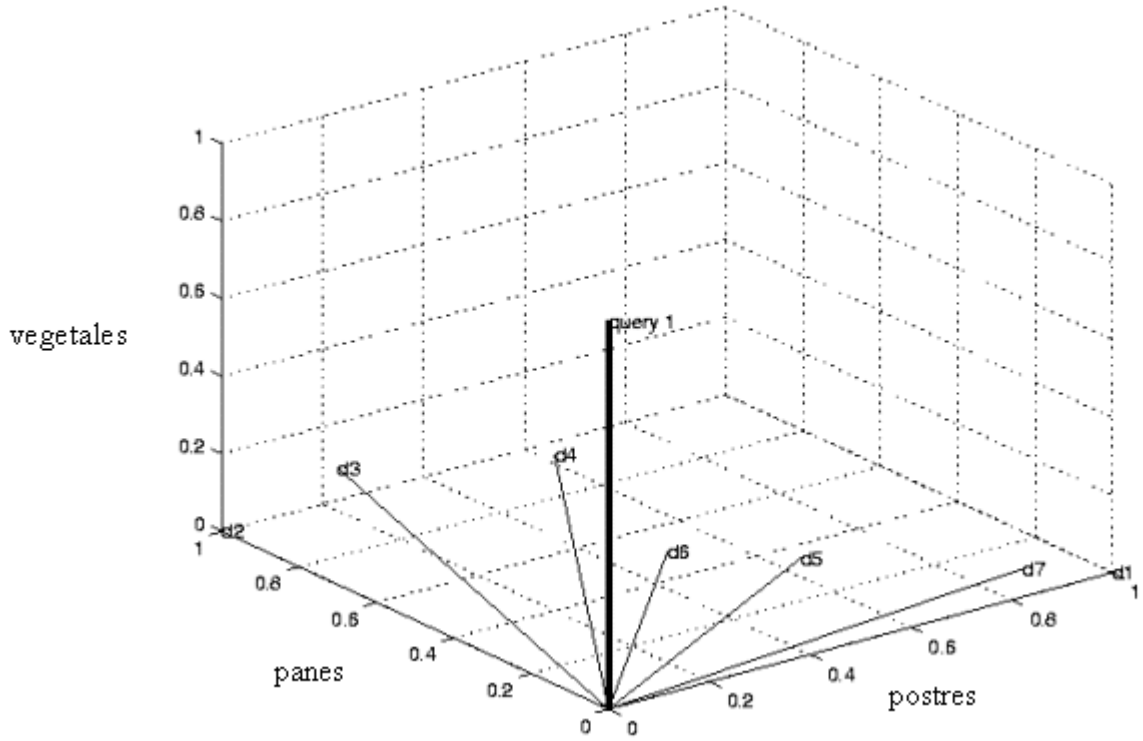
- De estos vectores se va a obtener una matriz *A*, donde cada columna es un documento del tema sobre Comida:

$$A = \begin{array}{c} \\ \textit{postres} \\ \textit{panes} \\ \textit{vegetales} \end{array} \begin{array}{c} \begin{array}{ccccccc} d_1 & d_2 & d_3 & d_4 & d_5 & d_6 & d_7 \end{array} \\ \begin{pmatrix} 0.146 & 0 & 0 & 0.146 & 0.146 & 0.146 & 0.029 \\ 0 & 0.067 & 0.067 & 0.067 & 0.033 & 0.067 & 0.067 \\ 0 & 0 & 0.544 & 0.544 & 0 & 0 & 0 \end{pmatrix} \end{array}$$


vegetales
1
0.8
0.6
0.4
0.2
0
d3
d4
d2
d6
d5
d7
d1
1
0.8
0.6
0.4
0.2
0
panes
0
0.2
0.4
0.6
0.8
1
postres

Ejemplo del Modelo Vectorial (cont.)

- Se quiere buscar los artículos relacionados con vegetales, entonces el vector consulta es $q_1 = (0,0,0.544)$.



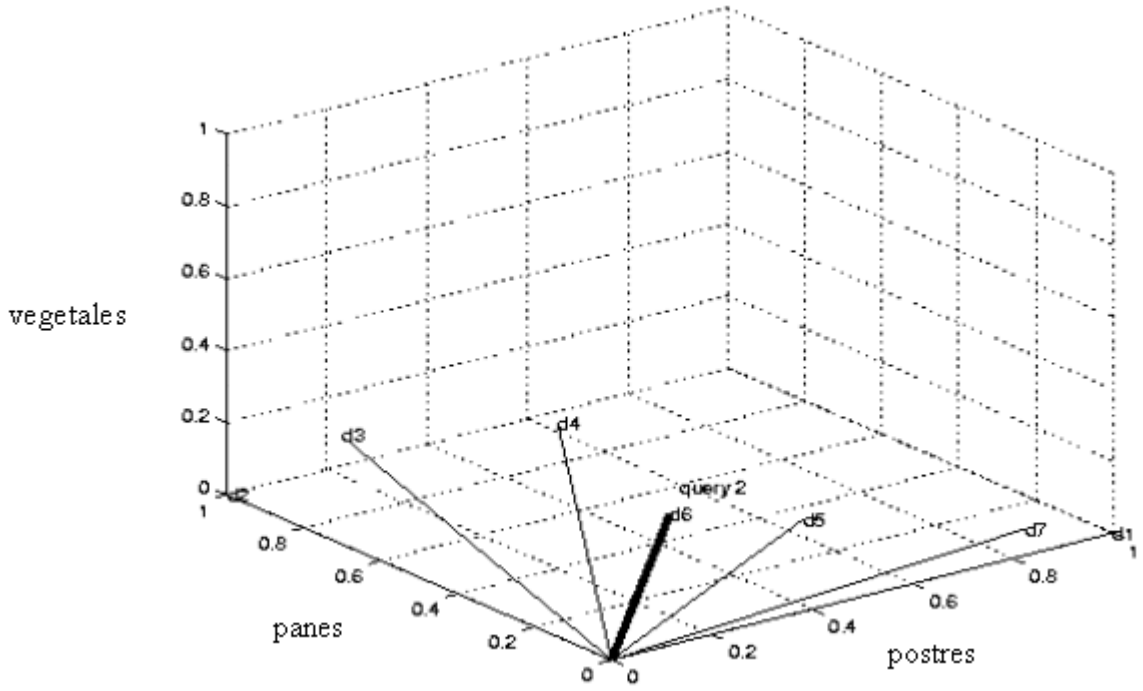
Ejemplo del Modelo Vectorial (cont.)

- La similaridad entre los vectores documentos y el vector consulta q_1 es:

Vector Documento	Coseno del Ángulo entre d_i y q_1
d1	0.000
d2	0.000
d3	0.993
d4	0.959
d5	0.000
d6	0.000
d7	0.000

Ejemplo del Modelo Vectorial (cont.)

- Se quiere buscar los artículos relacionados con postres y panes, entonces el vector consulta es q_2 = (0.146,0.067,0).



Ejemplo del Modelo Vectorial (cont.)

- La similaridad entre los vectores documentos y el vector consulta q_2 es:

Vector Documento	Coseno del Ángulo entre d_i y q_2
d1	0.909
d2	0.417
d3	0.051
d4	0.283
d5	0.979
d6	1.000
d7	0.745

Ventajas del Modelo Vectorial

- Se mejora el rendimiento con las fórmulas de obtención de pesos.
- Se pueden recuperar documentos que se “aproximen” a la consulta.
- La fórmula del coseno proporciona, además, un ranking sobre la respuesta.
- Es muy elástico como estrategia de ranking en colecciones generales.
- En comparación con otros modelos, es superior o igual en rendimiento a las alternativas.
- Es simple y rápido.

Desventajas del Modelo Vectorial

- Considera los términos como independientes, lo que puede causar bajo rendimiento (en teoría).
- La forma de pesado es intuitiva, pero no es muy formal.
- Es difícil de mejorar sin expansión de consultas o retroalimentación por relevancia (*relevance feedback*).

Modelo Probabilístico

Características del Modelo Probabilístico

- Los documentos y la consulta se representan como un conjunto de términos.
- También se le llama *binary independence retrieval model*.
- Se presupone que existe exactamente un subconjunto de documentos que son relevantes para una consulta dada.
- La idea del modelo es que dada una consulta, existe exactamente un conjunto de documentos, y no otro, que satisface dicha consulta. Este conjunto es el **conjunto ideal**.
- Por tanto, el problema de RI será el proceso de especificar las propiedades del conjunto ideal.

Características del Modelo Probabilístico (cont.)

- El problema es que no conocemos exactamente las propiedades del conjunto ideal.
- Deberemos realizar una suposición inicial sobre estas propiedades para tratar de refinarlas consulta tras consulta.
- Tras cada consulta, el usuario determinará los documentos que son relevantes, con lo que se podrá refinar la descripción del conjunto ideal.
- Su base teórica es la **Teoría de la Probabilidad**.
- Para cada documento, se intenta evaluar la *probabilidad* de que el usuario lo considere relevante.

Características del Modelo Probabilístico (cont.)

- Recupera los documentos que con mayor probabilidad son relevantes. Sin embargo, es bastante poco popular.
- Tiene una base teórica distinta a la del modelo vectorial y permite extensiones que sí son populares.

Caracterización Formal del Modelo Probabilístico

$$[\ D, Q, F, R(q_i,d_j)\ ]$$

- **D** = Conjunto de vectores que representan a los documentos, formado por los pesos (0 ó 1) de los términos de la colección en el documento.
- **Q** = Conjunto de vectores que representan a las consultas, formado por los pesos (0 ó 1) de los términos de la colección en la consulta.
- ***F*** = Teoría de la Probabilidad.
- $\boldsymbol{R(q_i,d_j)}$ = Principio de Probabilidad.

Nomenclatura

- $P(R)$ $\rightarrow$ probabilidad de que seleccionando cualquier documento de la colección sea relevante
- $P(R')$ $\rightarrow$ probabilidad de que seleccionando cualquier documento de la colección sea no relevante
- $P(R \mid d_j)$ $\rightarrow$ probabilidad de que el documento seleccionado sea relevante
- $P(R' \mid d_j)$ $\rightarrow$ probabilidad de que el documento seleccionado sea no relevante

Nomenclatura (cont.)

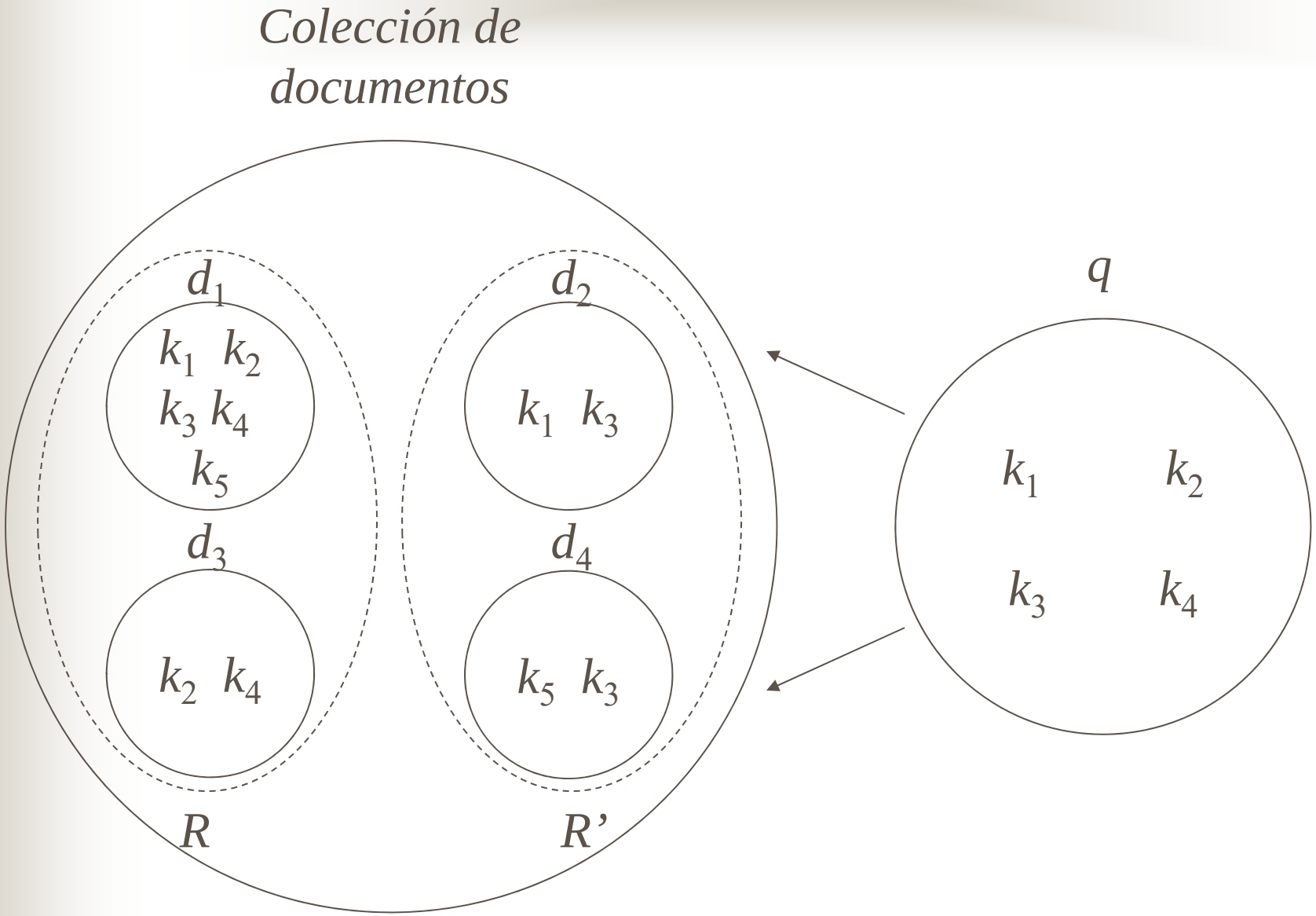
- $P(d_j \mid R)$ $\rightarrow$ probabilidad de elegir aleatoriamente el documento d_j entre los documentos relevantes
- $P(d_j \mid R')$ $\rightarrow$ probabilidad de elegir aleatoriamente el documento d_j entre los documentos no relevantes
- $P(k_i \mid R)$ $\rightarrow$ probabilidad de que el término aparezca en un documento relevante
- $P(k_i \mid R')$ $\rightarrow$ probabilidad de que el término aparezca en un documento no relevante

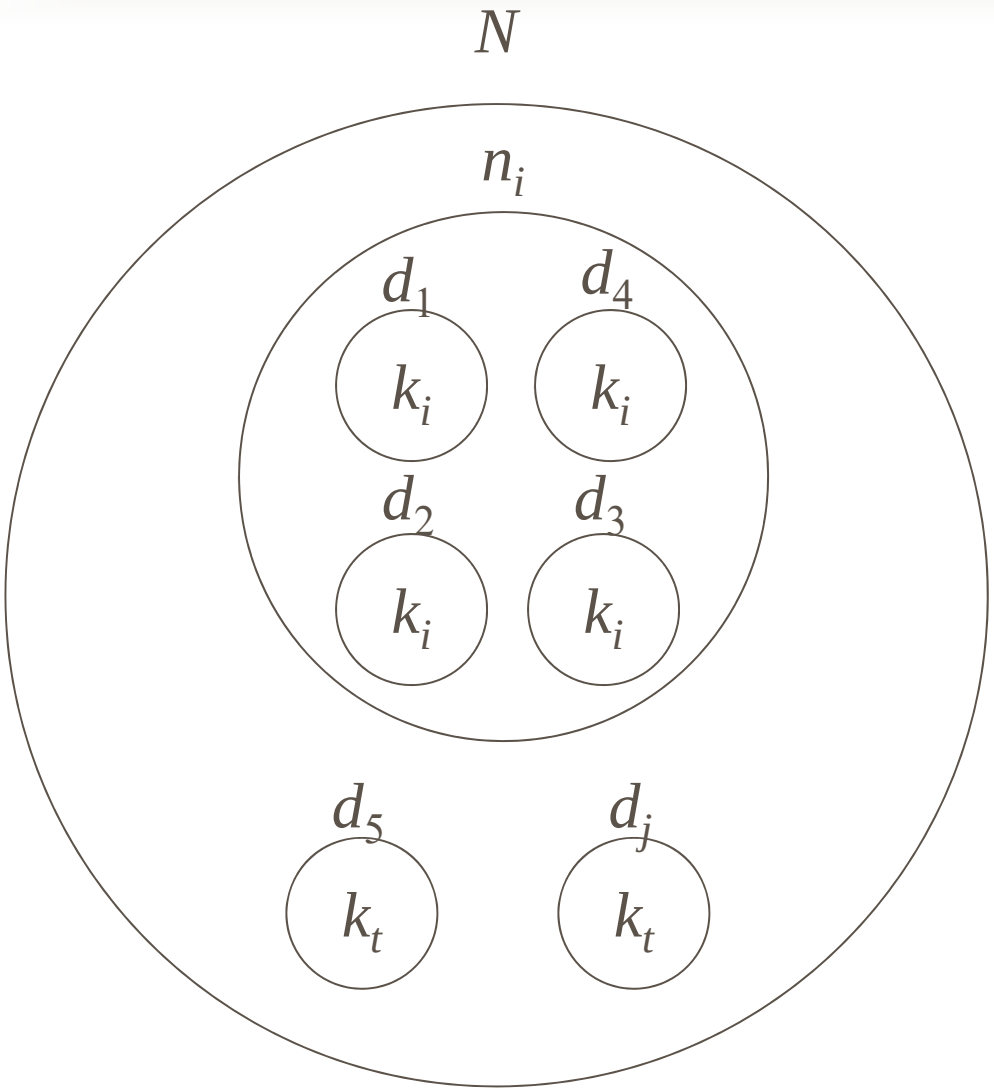
Nomenclatura (cont.)

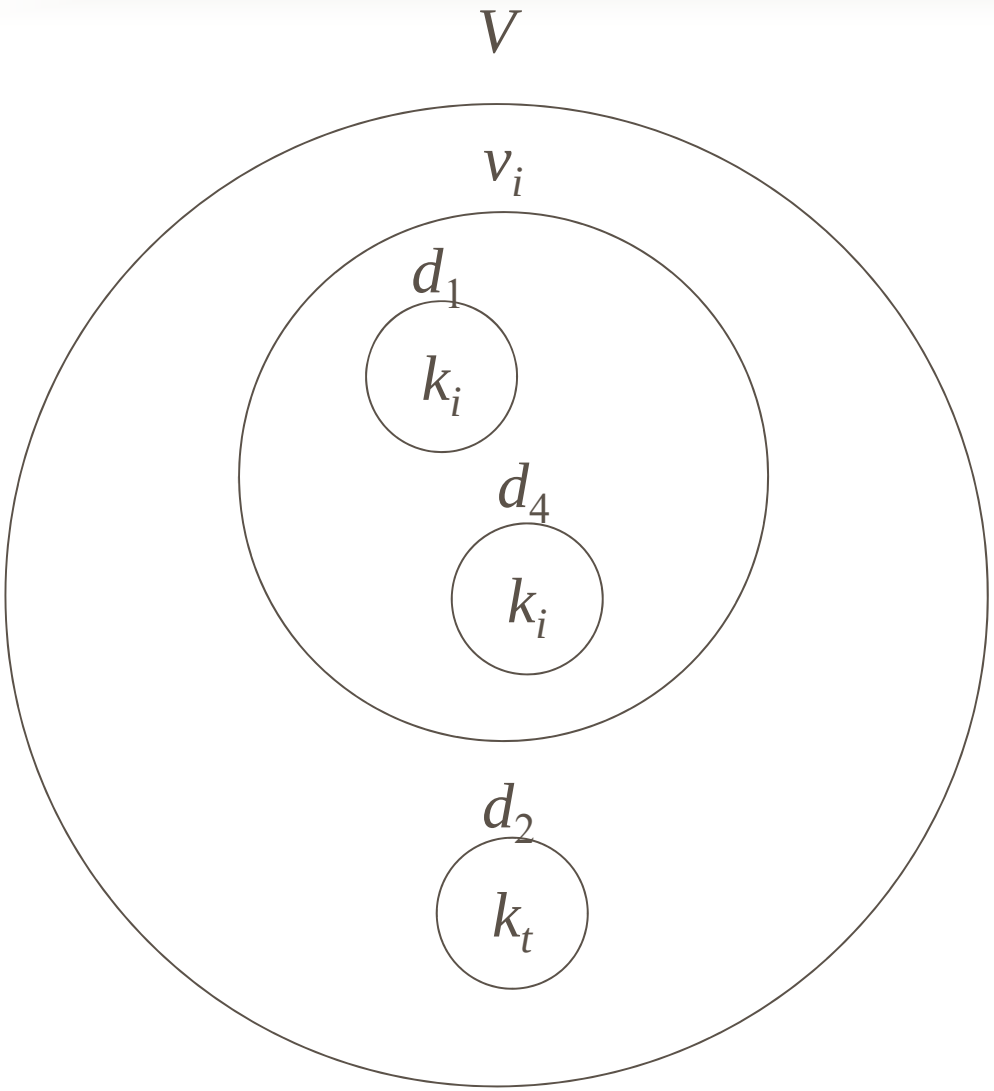
- $P(k_i' \mid R)$ → probabilidad de que el término no aparezca en un documento relevante
- $P(k_i' \mid R')$ → probabilidad de que el término no aparezca en un documento no relevante
- $P(d_j)$ → probabilidad de obtener el documento d_j aleatoriamente seleccionando uno de entre toda la colección.
- w_{iq} → peso de relevancia;
  i = término, q = consulta

Nomenclatura (cont.)

- $\prod g_i(d_j)=1 \rightarrow$ obtiene todos los términos del documento que tengan peso igual a 1
- $\prod g_i(d_j)=0 \rightarrow$ obtiene todos los términos del documento que tengan peso igual a 0
- $N \rightarrow$ total de documentos de una colección
- $n_i \rightarrow$ el número de documentos en los que aparece el término en el documento
- $V \rightarrow$ total de documentos de una colección recuperados
- $v_i \rightarrow$ el número documentos recuperados en los que aparece el término en el documento


Colección de documentos
d_1
k_1 k_2
k_3 k_4
k_5
d_2
k_1 k_3
d_3
k_2 k_4
d_4
k_5 k_3
R
R'
q
k_1 k_2
k_3 k_4


N
n_i
d_1
k_i
d_4
k_i
d_2
k_i
d_3
k_i
d_5
k_t
d_j
k_t


V
v_i
d_1
k_i
d_4
k_i
d_2
k_t

Principio de probabilidad

- Dada una consulta q y un documento d_j se trata de determinar la probabilidad de que el usuario encuentre el documento relevante.
- Se asume que esta probabilidad de relevancia depende sólo de las representaciones del documento y de la consulta.
- Se asume que hay un subconjunto de todos los documentos que el usuario prefiere como respuesta a su consulta, llamado **conjunto de respuesta ideal** y se denota por R.
- El conjunto R debería maximizar la probabilidad global de relevancia para el usuario. Los documentos que no pertenezcan al conjunto serán considerados como no relevantes para el usuario.

Principio de probabilidad (cont.)

- La relevancia de un documento se calcula como:

$$\frac{P(d_j \ relevante \ para \ q)}{P(d_j \ no \ relevante \ para \ q)}$$

Definición Formal del Modelo Probabilístico

- Los pesos de los términos índice son binarios ($w_{ij} \in \{0,1\}$, $w_{iq} \in \{0,1\}$). Una consulta q es un subconjunto de términos índice. Sea R el conjunto de documentos conocidos (o inicialmente supuestos) como relevantes. Sea R' el complemento de R. Sea $P(R|d_j)$ la probabilidad de que el documento d_j sea relevante a la consulta q y $P(R'|d_j)$ la probabilidad de que d_j no sea relevante a q. Entonces, la similitud del documento con la consulta se define como:

$$sim(d_j, q) = \frac{P(R \mid d_j)}{P(R' \mid d_j)} = \frac{\frac{P(d_j \mid R) \times P(R)}{P(d_j)}}{\frac{P(d_j \mid R') \times P(R')}{P(d_j)}} \approx \frac{P(d_j \mid R)}{P(d_j \mid R')}$$

Definición Formal del Modelo Probabilístico

- La similitud del documento con la consulta se define como:

$$sim(d_j, q) \approx \frac{\prod_{gi(d_j)=1} P(k_i \mid R) * \prod_{gi(d_j)=0} P(k_i' \mid R)}{\prod_{gi(d_j)=1} P(k_i \mid R') * \prod_{gi(d_j)=0} P(k_i' \mid R')}$$

$$sim(d_j, q) \approx \sum_{i=1}^{t} w_{iq} * w_{ij} * (\log \frac{P(k_i \mid R)}{1 - P(k_i \mid R)} + \log \frac{1 - P(k_i \mid R')}{P(k_i \mid R')})$$

Definición Formal del Modelo Probabilístico

- Inicialmente se supone que:

$$P(k_i \mid R) = 0.5 \qquad P(k_i \mid R') = \frac{n_i}{N}$$

- Luego de una iteración se recuperan V documentos; sea v_i el número de documentos recuperados que contienen el término k_i. Se recalcula:

$$P(k_i \mid R) = \frac{v_i + 0.5}{V + 1} \qquad P(k_i \mid R') = \frac{n_i - v_i + 0.5}{N - V + 1}$$

Definición Formal del Modelo Probabilístico (cont.)

- El cociente es ahora fácil de calcular con las probabilidades de que los términos del documento estén o no estén en los documentos de los conjuntos relevantes o no relevantes, según el caso.
- Para que quede claro, un documento será relevante si:

$$P(R|d_j) > P(R'|d_j)$$

ó

$$P(d_j|R) > P(d_j|R')$$

Ventajas del Modelo Probabilístico

- Los documentos se presentan en orden decreciente de probabilidad de relevancia.
- La relevancia de cada documento es independiente de la relevancia de otros.
- Tiene una base teórica distinta que permite extensiones que son muy populares.

Desventajas del Modelo Probabilístico

- No podemos calcular exactamente las probabilidades, y tenemos que hacer estimaciones.
- Hay que hacer una separación inicial de documentos en relevantes y no relevantes, se debe comenzar adivinando y luego refinar esa apuesta iterativamente.
- Es binario (no se consideran frecuencias de aparición de términos en los documentos), ve cada documento como un conjunto de términos.
- Se asume la independencia de términos, lo necesita.
- Existen estudios que muestran que es inferior al vectorial, y toda la comunidad científica lo considera inferior al vectorial.

Comparación de los Modelos Clásicos de RI

- El modelo booleano es el más flojo de todos los clásicos. No permite relevancias parciales y ofrece problemas de rendimiento.
- El modelo vectorial ofrece mejores resultados que el probabilístico, pero para colecciones generalistas.
- Ninguno establece relaciones entre los términos, todos los términos son tomados de forma independiente.

Evaluación de los Modelos Clásicos de RI

- El modelo clásico más popular es el vectorial, por ser simple, fácil y eficiente de implementar, y entregar buenos resultados. En muchos casos las aplicaciones llegan hasta aquí.
- Todos los modelos clásicos tienen ciertas falencias comunes, la más notoria es la incapacidad para capturar las relaciones entre términos.
- Por ejemplo: Si se busca sobre "*guerra fría*" se quisiera recuperar un documento que habla sobre "*la crisis de los misiles cubanos*". Sin embargo, el sistema no tiene idea de que ambas cosas están relacionadas y la intersección de vocabulario puede ser nula.

Evaluación de los Modelos Clásicos de RI (cont.)

- Parte de la solución pasa por el análisis lingüístico:
  - *Lematizar:* para no perderse variantes de la misma palabra.
  - *Etiquetar:* para distinguir verbos y sustantivos.
  - *Detectar frases comunes:* para no recuperar información sobre heladeras cuando se pregunte por "*guerra fría*".
  - Es posible hacer un análisis lingüístico más fino, pero la tecnología existente tiene sus limitaciones.
- Otro elemento es el uso de *tesauros* (sinónimos) para expandir la consulta, de modo que se pueda recuperar "*vendo camioneta usada*" frente a la consulta "*autos de segunda mano*". Sin embargo, mantener un buen tesauro es costoso en términos de trabajo manual y no suele ser posible mantenerlo al día con las expresiones que van apareciendo.

Evaluación de los Modelos Clásicos de RI (cont.)

- Además, un tesauro global no funciona siempre bien, por ejemplo: *"estrella"* puede ser una generalización de *"supernova"*, pero no en un contexto que habla de estrellas de cine y televisión.
- Se puede utilizar información del texto como la estructura para refinar mejor la consulta (por ejemplo: un término que aparece en el título debe ser más relevante que en un pie de página).
- Se pueden utilizar distintas técnicas para descubrir que ciertos términos están correlacionados. A esto apuntan los modelos alternativos y las técnicas de expansión de consultas.

Referencias Bibliográficas

- La información fue tomada de:
  - Libro de texto del curso.