Análisis y Recuperación de Información

1er Cuatrimestre 2017

Página Web

http://www.exa.unicen.edu.ar/catedras/ayrdatos/

Prof. Dra. Daniela Godoy

ISISTAN Research Institute
UNICEN University
Tandil, Bs. As., Argentina
http://www.exa.unicen.edu.ar/~dgodoy
dgodoy@exa.unicen.edu.ar

Modelos de IR

- Un modelo de IR es la representación abstracta de un proceso
  - los modelos permiten estudiar propiedades, sacar conclusiones y hacer predicciones
  - la calidad de las conclusiones dependerá de qué tanto el modelo se ajuste a la realidad
- Un modelo de IR describe los procesos humanos y computacionales involucrados en la recuperación
  - el comportamiento de una persona que intenta recuperar información
  - la forma en que se rankean los documentos
  - los componentes del sistema, como usuarios, necesidades de información, consultas, documentos, cálculo de relevancia, etc.

Modelos de IR

- Un modelo de IR especifica:
  - la representación de documentos
  - la representación de consultas
  - la función de recuperación
- También determina la noción de relevancia, binaria o continua

Modelos de IR

- Matching exacto:
  - la consulta especifica un criterio de recuperación preciso
  - cada documento coincide o no con la consulta
  - el resultado es un conjunto de documentos, usualmente sin orden
- Matching aproximado:
  - la consulta describe un criterio de recuperación de los documentos deseados
  - cada documento tiene un grado de coincidencia con la consulta
  - el resultado es una lista ordenada de documentos, el primero es el “mejor”

Modelos de IR

- Modelo Booleano (teoría de conjuntos)
  - Booleano extendido
- Modelo Probabilístico (teoría de probabilidades)
- Modelo de Espacio de Vectores (algebraico/estadístico)
  - Espacio de vectores generalizado

Modelos de IR

- Cada documento está representado por un conjunto de keywords o términos indexados
- Un término indexado es una palabra que es útil para recordar el contenido o tema de un documento
- Los términos indexados pueden ser los sustantivos, que poseen un significado asociado
  - si se usan solo los sustantivos se reduce el tamaño del índice, pero requiere identificarlos → Part of Speech tagger
- Los buscadores asumen que todas las palabras son términos indexables (full text representation)

Modelos de IR

- No todos los términos son igualmente útiles para representar el contenido de un documento
- La importancia de los términos indexados está representada por el peso que se asocia a ellos:
  - k_i *es un término indexados*
  - d_j *es un documento*
  - w_{ij} *es el peso asociado con* k_i *en* d_j
- El peso cuantifica la importancia del término para describir el contenido de un documento

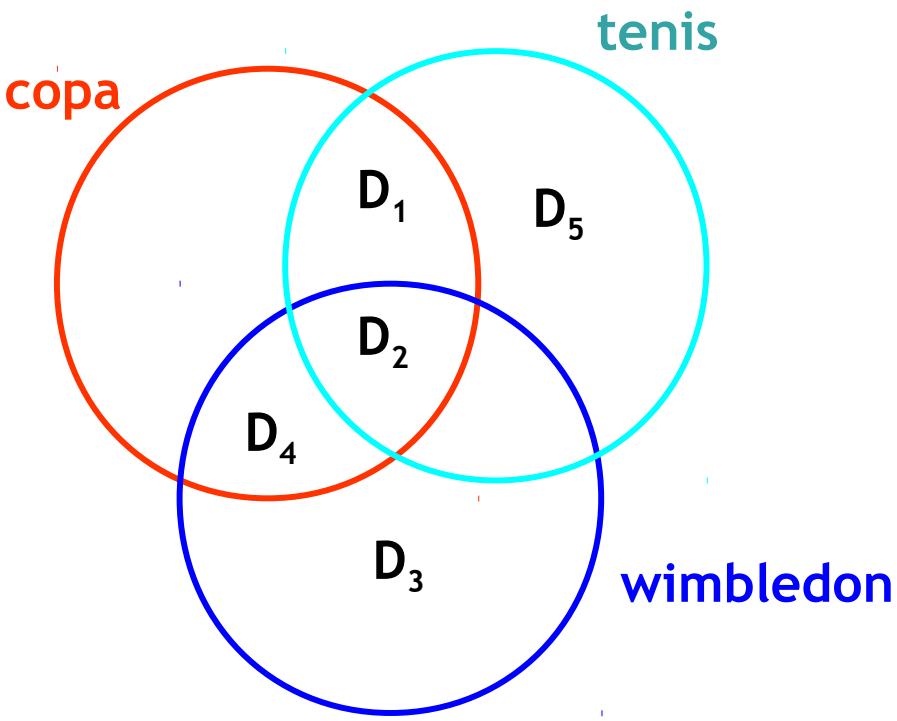
Modelo Booleano

- Modelo de matching exacto basado en la teoría de conjuntos
- Las consultas son expresiones booleanas, los documentos son recuperados si satisfacen tal expresión booleana
- Un términos está presente o ausente en un documento, no hay grados de pertenencia
- Los documentos se recuperan sin un orden en particular
- Fue el principal modelo por más de tres decadas, muchos sistemas de búsqueda que usan hoy en día son booleanos (por ejemplo, mail, catalogo de librerías, Mac OS X Spotlight)

Modelo Booleano

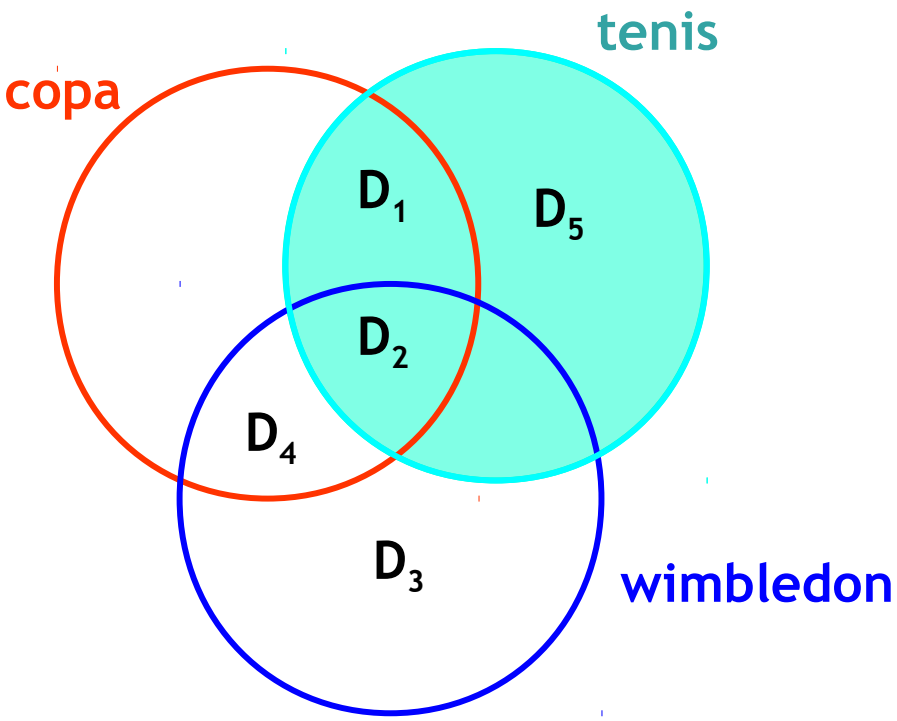
- Un documento se representa mediante una conjunción de términos
  - D_1=(tenis AND copa AND davis)
  - D_2=(tenis AND wimblendon AND roland AND garros)
- Una consulta es una combinación de términos unidos por los conectores AND, OR y NOT
  - Q_1=(tenis OR fútbol)
- La función de comparación recupera un documento D ante una consulta Q sssi Q es una consecuencia lógica de D

Modelo Booleano



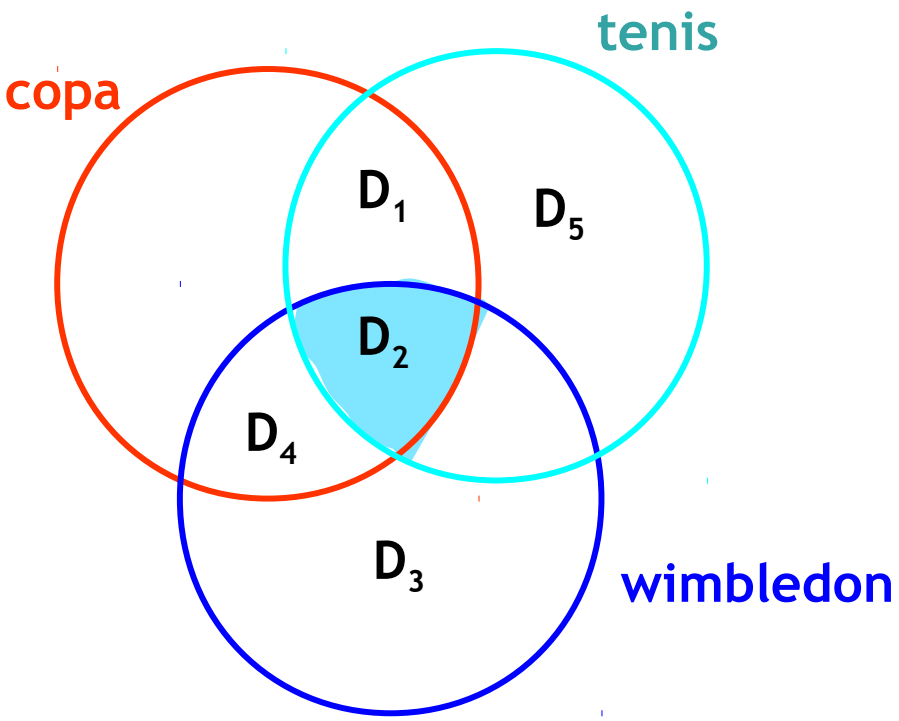
- Documentos:
  - D_1=(tenis AND copa)
  - D_2=(tenis AND wimbledon AND copa))
  - D_3=(wimbledon)
  - D_4=(copa AND wimbledon)
  - D_5=(tenis)

Modelo Booleano



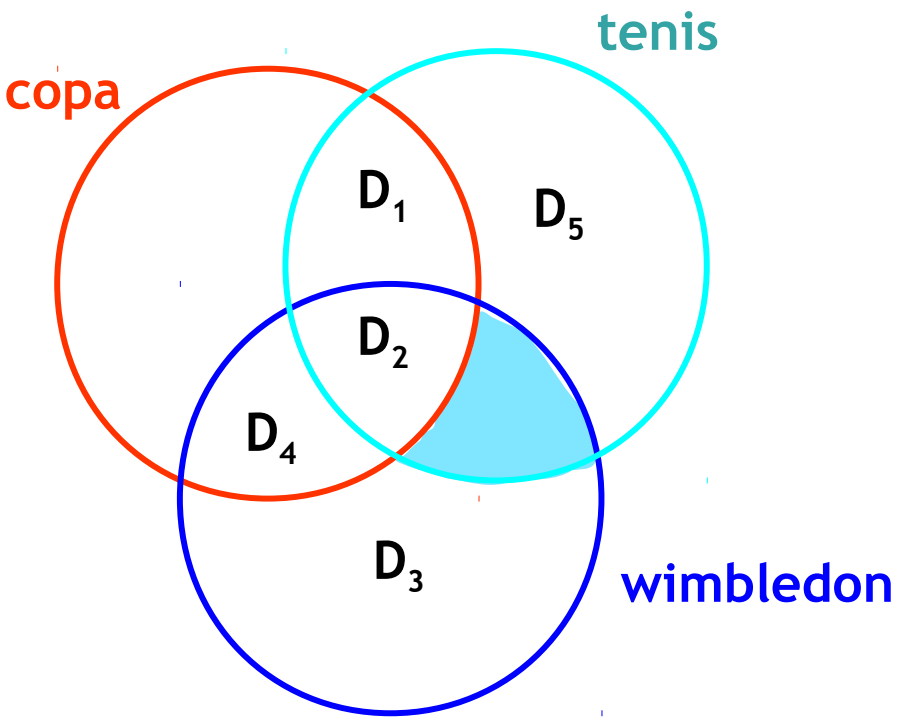
- Documentos:
  - D_1=(tenis AND copa)
  - D_2=(tenis AND wimbledon AND copa))
  - D_3=(wimbledon)
  - D_4=(copa AND wimbledon)
  - D_5=(tenis)
- Consulta
  - Q_1=(tenis) ➔ D_1, D_2, D_5

Modelo Booleano



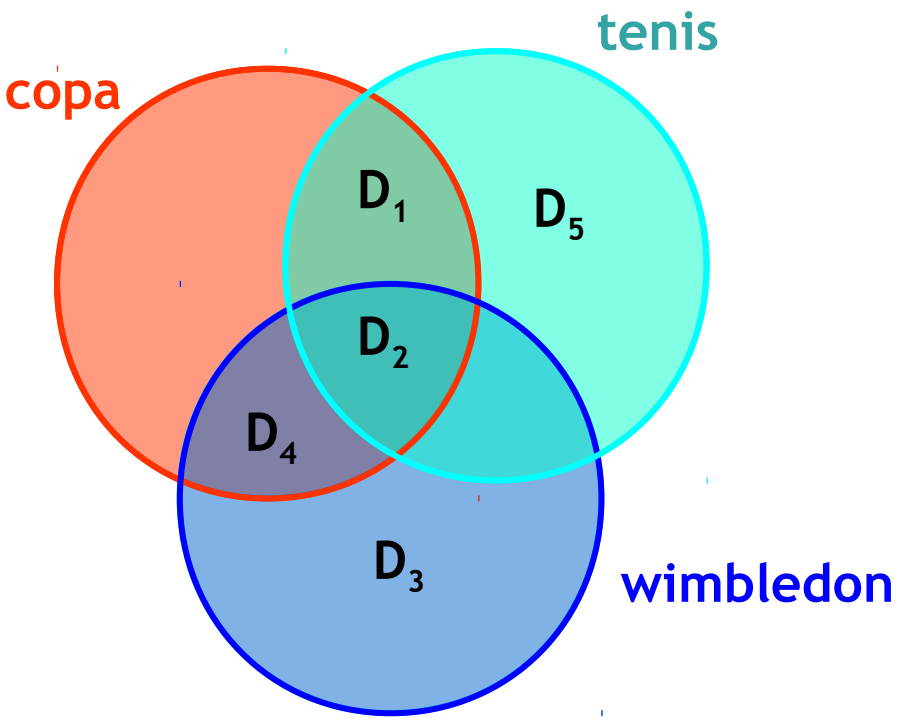
- Documentos:
  - D_1=(tenis AND copa)
  - D_2=(tenis AND wimbledon AND copa))
  - D_3=(wimbledon)
  - D_4=(copa AND wimbledon)
  - D_5=(tenis)
- Consulta
  - Q_1=(tenis) ➔ D_1, D_2, D_5
  - Q_2=(tenis AND wimbledon AND copa) ➔ D_2

Modelo Booleano



- Documentos:
  - D_1=(tenis AND copa)
  - D_2=(tenis AND wimbledon AND copa))
  - D_3=(wimbledon)
  - D_4=(copa AND wimbledon)
  - D_5=(tenis)
- Consulta
  - Q_1=(tenis) ➔ D_1, D_2, D_5
  - Q_2=(tenis AND wimbledon AND copa) ➔ D_2
  - Q_3=(tenis AND wimbledon AND NOT copa) ➔

Modelo Booleano



- Documentos:
  - D_1=(tenis AND copa)
  - D_2=(tenis AND wimbledon AND copa))
  - D_3=(wimbledon)
  - D_4=(copa AND wimbledon)
  - D_5=(tenis)
- Consulta
  - Q_4=(tenis OR wimbledon OR copa) ➔ D_1, D_2, D_3, D_4, D_5

Modelo Booleano

- Ventajas
  - consultas simples y fáciles de entender
  - implementación relativamente sencilla
- Desventajas
  - recuperación esta basada en decisiones binarias sin noción de matching parcial
  - el modelo booleano puro no ofrece rankings de documentos, en la práctica se usa:
    - orden cronológico
    - orden por número total de *hits* sobre los términos consultados
  - la formulación de consultas en forma de expresiones lógicas puede ser dificultosa para usuarios inexpertos
  - las consultas devuelven muchos documentos o muy pocos para una consulta del usuario

Modelo Booleano

- Desventajas
  - impone un criterio binario para decidir la relevancia
  - el problema de extender el modelo Booleano para matching parcial y ranking ha recibido mucha atención
  - posibles extensión es el modelo de conjuntos difusos

Modelo Probabilístico

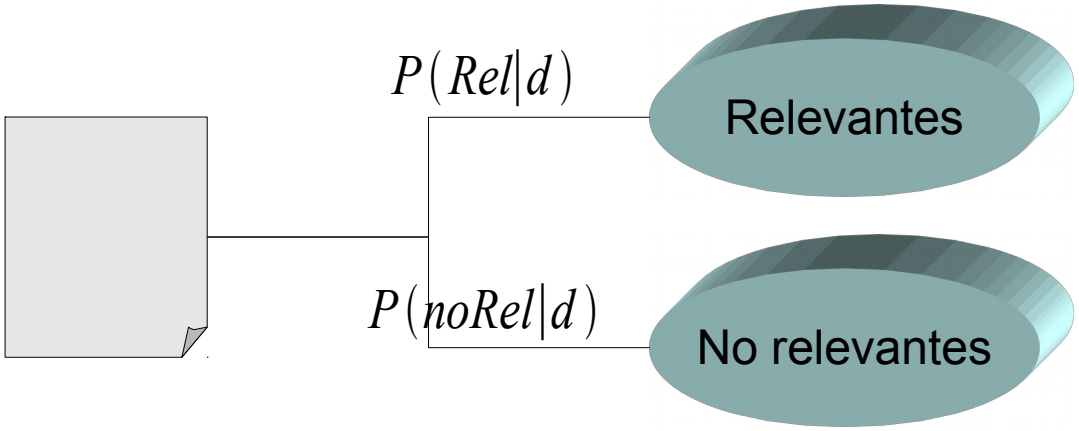
- Captura el problema de IR en un marco probabilístico
- Dada una consulta *q* y un documento *d* en la colección, el modelo estima la probabilidad de que el usuario evalúe el documento *d* como relevante
- Intenta responder a la pregunta: cuál es la probabilidad de que un documento sea relevante a una consulta dada?

$$P(Rel|d)$$

probabilidad de relevancia dado el documento *d*

Modelo Probabilístico

- Dos tipos de respuesta a una consulta: documentos relevantes e irrelevantes
- Asume que existe un subconjunto *R* de la colección que contiene sólo los documentos relevantes
- La respuesta ideal debería ser *R* que maximiza la probabilidad de relevancia



Modelo Probabilístico

- Se recupera un conjunto de documentos inicial con algún otro método (booleano o vectorial)
- El usuario inspecciona los documentos buscando aquellos relevantes (sólo los primeros 10 o 20)
- El sistema de IR usa esta información para refinar la descripción del conjunto ideal y se repite el proceso para mejorar tal descripción
- La descripción del conjunto ideal se modela en términos probabilísticos

Modelo Probabilístico

- Un documento se recupera si la probabilidad de pertenecer al conjunto de documentos relevantes es mayor que la de pertenecer a los no relevantes:

$$P(Rel|d) > P(noRel|d)$$

- La similitud de un documento a una consulta:

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

Es igual para todos los documentos

Teorema de Bayes

$$\frac{P(d|Rel)\,P(Rel)}{P(d|noRel)\,P(noRel)} \sim \frac{P(d|Rel)}{P(d|noRel)}$$

P(Rel) es la probabilidad de que un documento elegido aleatoriamente sea relevante a consulta
P(d|Rel) es la probabilidad de seleccionar aleatoriamente el documento *d* del conjunto *Rel*

Modelo Probabilístico

- Asumiendo independencia entre lo términos:

$$P(d|Rel) = \prod P(a_i|Rel)$$

  - la probabilidad de que un documento sea relevante se basa en la probabilidad de relevancia de los términos individuales

$$P(a_i|Rel)$$

  - la probabilidad de que a_i esté presente en un documento relevante

$$P(a_i|noRel)$$

  - la probabilidad de que a_i esté presente en un documento no relevante

Modelo Probabilístico

- Estimar la P(a_i|Rel) y P(a_i|noRel):
  - Valores constantes iniciales

$$P(a_i|Rel)=0.5$$

$$P(a_i|noRel)=\frac{n_i}{N}$$

n_i es el número de documentos que contiene a_i

  - Proceso iterativo para mejorar los valores iniciales:

$$P(a_i|Rel)=\frac{V_i}{V}$$

$$P(a_i|noRel)=\frac{n_i-V_i}{N-V}$$

V es el conjunto de documentos recuperados

V_i es el conjunto de documentos en V que contienen a_i

Modelo Probabilístico

- Ventajas
  - los documentos se rankean en base a la probabilidad de ser relevantes
- Desventajas
  - la necesidad de una separación inicial de los documentos en relevantes e irrelevantes
  - no toma en cuenta la frecuencia de los términos
  - asume independencia entre las palabras

Modelo de Espacio de Vectores

- Modelo de matching aproximado
- Asume que cualquier objeto textual puede representarse mediante un vector de términos
  - Ejemplo: documentos, consultas, etc.
- La similitud se determina mediante la distancia en el espacio de vectores

Modelo de Espacio de Vectores

- Después de procesar los documentos quedan *t* términos distintos
  - Términos únicos que forman el VOCABULARIO
- Estos términos forman un espacio de vectores
  - Dimensión = *t* = |vocabulario|
  - 2 términos→ bi-dimensional; …; *n*-términos → *n*-dimensional
- Cada término *i* en un documento d_j o consulta q_j tiene asociado un peso w_{ij}
- Los documentos y las consultas se expresan como vectores *t*-dimensionales:

  $d_j = (w_{1j}, w_{2j}, \ldots, w_{tj})$

  $q_j = (w_{1j}, w_{2j}, \ldots, w_{tj})$

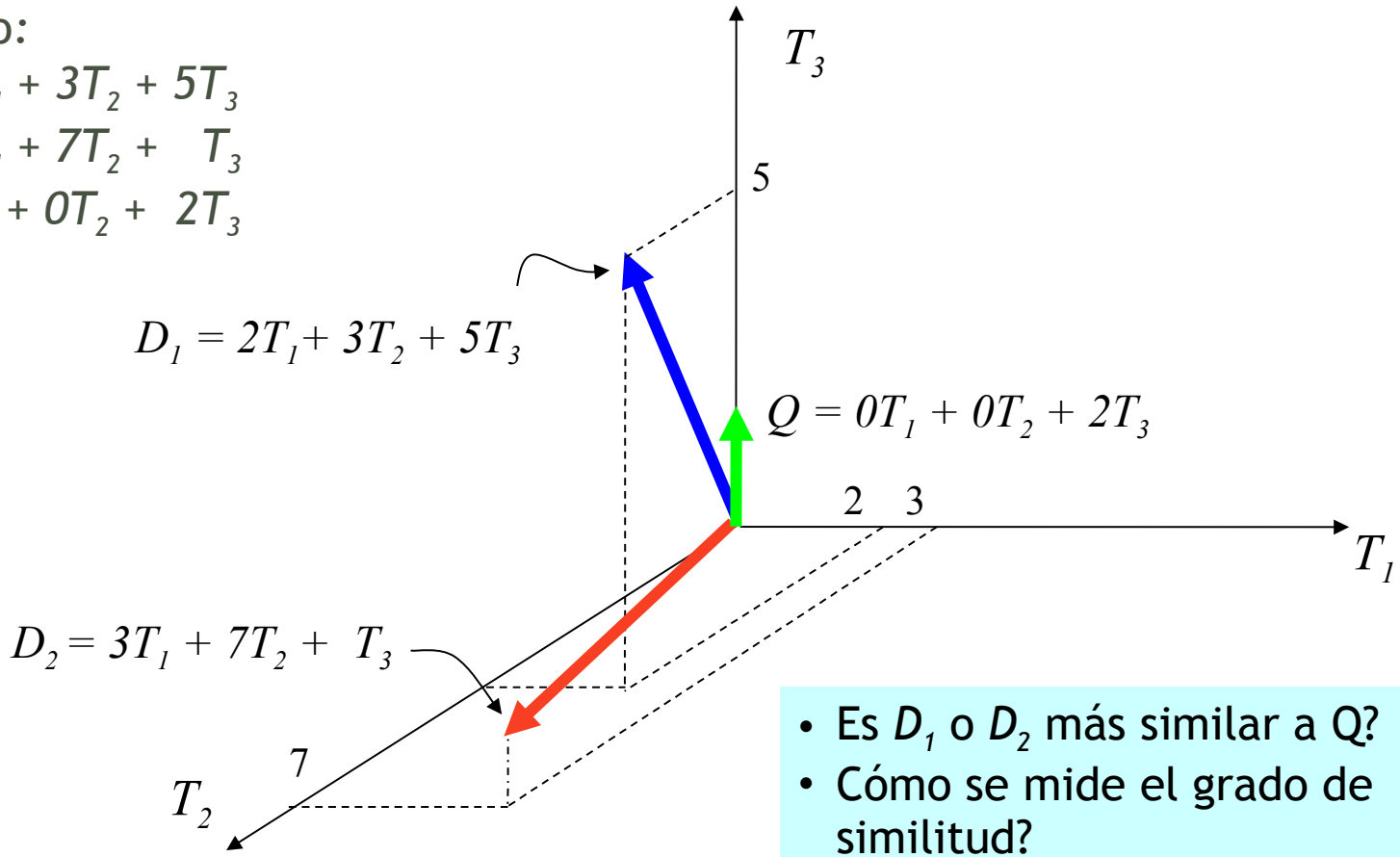
Modelo de Espacio de Vectores

Ejemplo:

$D_1 = 2T_1 + 3T_2 + 5T_3$

$D_2 = 3T_1 + 7T_2 + T_3$

$Q = 0T_1 + 0T_2 + 2T_3$



- Es D_1 o D_2 más similar a Q?
- Cómo se mide el grado de similitud?

Modelo de Espacio de Vectores

- Una medida de similitud es una función que permite calcular el grado de cercanía de dos vectores en el espacio
- Usar una medida de similitud entre una consulta y un documento permite:
  - Ordenar los documentos recuperados de acuerdo a su relevancia
  - Controlar el número de documentos recuperados mediante el uso de un umbral

Modelo de Espacio de Vectores

- La similitud entre un documento D y una consulta Q puede calcularse

$$\text{sim}(d_j, q) = d_j \cdot q = \sum_{i=1}^{t} w_{ij} * w_{iq}$$

  - donde w_{ij} es el peso del término *i* en el documento *j* y w_{iq} es el peso del término *i* en la consulta
  - Para vectores binarios el producto interno es el número de términos que coinciden entre documentos y consultas (tamaño de la intersección)
  - Para vectores de numéricos, es la suma de los productos de los pesos de los términos coincidentes
- Un documento se recupera aún cuando coincida sólo parcialmente con los términos de la consulta

Modelo de Espacio de Vectores

Pesos binarios:

	retrieval	database	architecture	computer	text	management	information
D =	1,	1,	1,	0,	1,	1,	0
Q =	1,	0,	1,	0,	0,	1,	1

$sim(D, Q) = 3$

Tamaño del vector = vocabulario = 7

- Pesos no binarios:

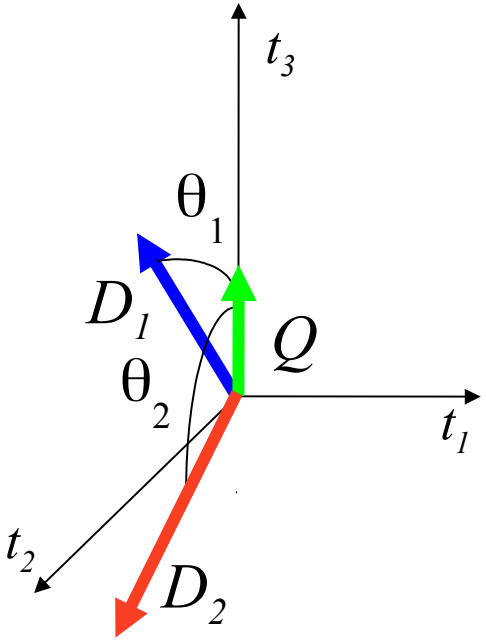
$D_1 = 2T_1 + 3T_2 + 5T_3$
$D_2 = 3T_1 + 7T_2 + 1T_3$
$Q = 0T_1 + 0T_2 + 2T_3$

$\text{sim}(D_1, Q) = 2*0 + 3*0 + 5*2 = 10$
$\text{sim}(D_2, Q) = 3*0 + 7*0 + 1*2 = 2$

Modelo de Espacio de Vectores

- La similitud del coseno mide el coseno del ángulo entre dos vectores
- Se calcula como el producto interno normalizado por la longitud de los vectores

$$\text{sim}(d_i, q) = \frac{d_i \cdot q}{|d_i||q|} = \frac{\sum_{k=1}^{n} w_{ki} * w_{kq}}{\sqrt{\sum_{k=1}^{n} w_{ki}^2} \sqrt{\sum_{k=1}^{n} w_{kq}^2}}$$



$D_1 = 2T_1 + 3T_2 + 5T_3$ $\quad \text{sim}(D_1, Q) = 10 / \sqrt{(4+9+25)(0+0+4)} = 0.81$

$D_2 = 3T_1 + 7T_2 + 1T_3$ $\quad \text{sim}(D_2, Q) = 2 / \sqrt{(9+49+1)(0+0+4)} = 0.13$

$Q = 0T_1 + 0T_2 + 2T_3$

D_1 es mucho más cercano que D_2 usando la medida de similitud del coseno pero la diferencia es menor usando el producto interno

Modelo de Espacio de Vectores

- Ventajas:
  - enfoque simple basado en nociones algebraicas
  - provee matching parcial y resultados rankeados
  - facilita una implementación eficiente para grandes colecciones de documentos
- Desventajas:
  - pérdida de información semántica (significado de las palabras)
  - pérdida de información sintáctica (frases, orden, etc.)
  - no tiene el control de un sistema booleano (ejemplo, si se requiere que un término determinado esté en un documento)