Modelos Clásicos de RI

Índice

- Introducción
- Modelo booleano
- Modelo vectorial
- Modelo probabilístico
- Modelos alternativos
- Conclusiones

Introducción

Introducción

- Los modelos de RI se centran fundamentalmente en el texto
  - Utilizan el texto de los documentos para devolver un ranking de ellos mismos de acuerdo a una consulta dada (textual)
- La web ha influenciado los modelos de RI
  - Es necesario tener en cuenta el concepto de enlace
- Existen otro tipo de información no textual que también debe indexarse
  - Videos, imágenes → Multimedia
- Tipos de modelos de RI
  - Texto
  - Enlaces
  - Multimedia

Introducción

- Texto
  - Modelos clásicos (texto no estructurado)
    - Boolean (teoría de conjuntos)
      - Fuzzy
      - Boolean extendido
      - Basado en conjuntos
    - Vector (algebraicos)
      - Modelo vectorial
      - Lantent Semantic Indexing
      - Redes Neuronales
    - Probabilístico
      - BM25
      - Modelos del lenguaje
      - Divergencia aleatoria
      - Redes bayesianas

Introducción

- Texto semi-estructurado
  - Proximidad de nodos
  - Basado en XML
- Enlaces (web)
  - Page Rank
  - Hub & Authorities
- Multimedia
  - Imágenes
  - Audio y música
  - Video

Introducción

- Modelos clásicos
  - Booleano
    - Los documentos y las consultas se representan como un conjunto de términos de indexación
    - Modelo basado en conjuntos
  - Vectorial
    - Los documentos y las consultas se representan como un vector $t - dimensional$
    - Modelo algebraico
  - Probabilístico
    - Basado en la teoría de la probabilidad

Modelo Booleano

Modelo Booleano

- Modelo basado en:
  - Teoría de conjuntos
  - Álgebra booleana
- Es un modelo sencillo con una semántica clara.
- Muchos de los primeros sistemas de Recuperación de información se basaban en este modelo.
- Características del modelo:
  - Los términos de indexación están presentes o ausentes en los documentos.
    - La matriz términos-documentos es binaria
  - La consulta es una expresión booleana.
    - Se pueden emplear tres tipos de conectivas: NOT, AND, OR

Modelo Booleano

- **Definición**. *En el modelo Booleano, todos los elementos de la matriz términos-documentos son 1, para indicar la presencia de un término en el documento, o 0, para indicar la ausencia de un término en el documento. Una consulta q es una expresión booleana sobre los términos de indexación (por ejemplo, $q = k_a \wedge (k_b \vee \neg k_c)$). Dada dicha consulta, la componente conjuntiva de términos que satisfaga las condiciones se llamará consulta componente conjuntiva $c(q)$. Utilizando todos los componentes de las consultas conjuntivas, la consulta se podrá expresar como una disyunción de todos los componentes. A esta consulta la llamaremos consulta en forma normal disyuntiva $QDNF$.*

Modelo Booleano

- Ejemplo
  - Consulta $\mathrm{q} = \mathrm{k_a} \wedge (\mathrm{k_b} \vee \neg \mathrm{k_c})$
  - Vocabulario = $\{\mathrm{k_a}, \mathrm{k_b}, \mathrm{k_c}\}$
  - $QDNF = (1,1,1) \vee (1,1,0) \vee (1,0,0)$
  - Supongamos un documento d_j que contiene los términos $\mathrm{k_a}$ y $\mathrm{k_c}$ pero no k_b.
    - $c(\mathrm{d_j}) = (1,0,1)$
    - No pertenece al $QDNF$
  - Por lo tanto, podemos decir que el documento $\mathrm{d_j}$ no satisface la consulta q

Modelo Booleano

- Ejemplo 2
  - Consulta $q = k_a \wedge (k_b \vee \neg k_c)$
  - Vocabulario = $\{k_a, k_b, k_c, k_d\}$
  - $QDNF = (1,1,1,0) \vee (1,1,1,1) \vee (1,1,0,0) \vee (1,1,0,1) \vee (1,0,0,0) \vee (1,0,0,1)$
  - Supongamos un documento d_j que contiene los términos k_a, k_b y k_c.
    - $c(d_j) = (1,1,1,0)$
    - Pertenece al $QDNF$
  - Por lo tanto, podemos decir que el documento d_j satisface la consulta q

Modelo Booleano

- **Definición**. *En el modelo booleano, una consulta puede definirse como una expresión booleana de términos de indexación. Sea $c(q)$ cualquier consulta expresada de forma conjuntiva. Dado un documento, d_j, sea $c(d_j)$ la correspondiente componente conjuntiva del documento. Entonces, la similitud del documento d_j con la consulta q puede definirse como:*

$$sim(d_j, q) = \begin{cases} 1 & Si\ \exists\ c(q) |\ c(q) = c(d_j) \\ 0 & en\ cualquier\ otro\ caso \end{cases}$$

- *Si $sim(d_j, q) = 1$, entonces según el modelo booleano, el documento d_j será relevante para la consulta q. En caso contrario, el documento no será relevante.*

Modelo Booleano

- Ventajas:
  - Definición formal clara
  - Simplicidad (peso binario de los términos)
- Desventajas:
  - No existe ranking
  - La formulación de consultas booleanas es muy compleja

Modelo Vectorial

Modelo Vectorial

- El modelo booleano es demasiado limitado.
  - No se permite una combinación parcial.
  - Los documentos o son relevante o no lo son.
    - No se puede establecer un ranking
  - Es difícil traducir una necesidad de información en una expresión booleana.
- Solución:
  - Asignar pesos no binarios a los términos de las consultas y documentos que serán utilizados para medir el grado de similitud entre los documentos del corpus y la consulta dada por el usuario.
  - Ordenar los documentos por similitud decreciente para obtener un ranking.

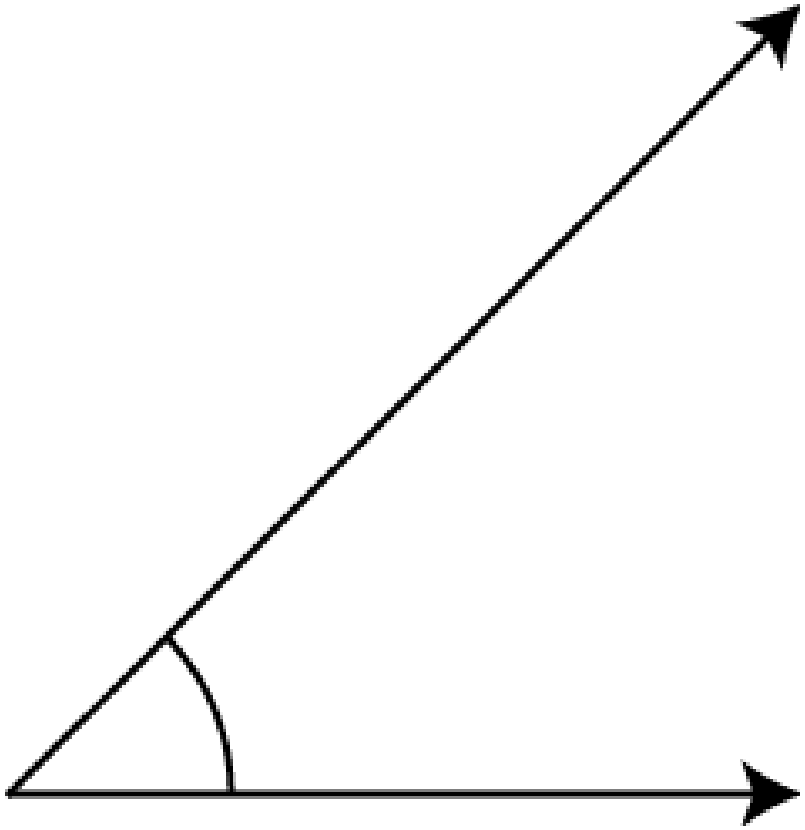
Modelo Vectorial

- **Definición**. En el modelo vectorial, el peso $w_{i,j}$ asociado con la pareja término-documento (k_i, d_j) es positivo y no binario. Asumimos que los términos de indexación son mutuamente independientes y son representados como vector unitarios en un espacio t-dimensional, donde t, es el número total de términos de indexación. En este sentido, la representación de un documento d_j y una consulta q puede expresarse de la siguiente forma:

$$\overrightarrow{d_j} = \left(w_{1,j}, w_{2,j}, \dots, w_{t,j}\right)$$
$$\vec{q} = \left(w_{1,q}, w_{2,q}, \dots, w_{t,q}\right)$$

- donde $w_{i,q}$ es el peso asociado con la pareja término-consulta (k_i, q), con $w_{i,q} \geq 0$

Modelo Vectorial



Modelo Vectorial

- La similitud entre un documento y una consulta puede calcularse mediante el coseno del ángulo entre los vectores que representan al documento d_j y a la consulta q.

$$\cos\theta = sim(d_j, q) = \frac{\overrightarrow{d_j} \cdot \vec{q}}{|\overrightarrow{d_j}| \times |\vec{q}|} = \frac{\sum_{i=1}^{t} w_{i,j} \times w_{i,q}}{\sqrt{\sum_{i=1}^{t} w_{i,j}^2} \times \sqrt{\sum_{i=1}^{t} w_{i,q}^2}}$$

- Donde $|\overrightarrow{d_j}|$ y $|\vec{q}|$ es el módulo de los vectores de la consulta y documento, y $\overrightarrow{d_j} \cdot \vec{q}$ es el producto de ambos vectores.
- Para la consulta, su peso se reduce al IDF ya que su frecuencia será 1.

Modelo Vectorial

- Ejemplo. Consulta “to do”.

documento	Cálculo del ranking	Ranking
d_1	$\frac{1 \times 3 + 0{,}415 \times 0{,}830}{5{,}068}$	0,660
d_2	$\frac{1 \times 2 + 0{,}415 \times 0}{4{,}899}$	0,408
d_3	$\frac{1 \times 0 + 0{,}415 \times 1{,}073}{3{,}762}$	0,118
d_4	$\frac{1 \times 0 + 0{,}415 \times 1{,}073}{7{,}738}$	0,058

Modelo Vectorial

#	Término	IDF	d_1	d_2	d_3	d_4
1	to	1	3	2		
2	do	0,415	0,830		1,073	1,073
3	is	2	4			
4	be	0				
5	or	2		2		
6	not	2		2		
7	I	1		2	2	
8	am	1		2	1	
9	what	2		2		
10	think	2			2	
11	therefore	2			2	
12	da	2				5,170
13	let	2				4
14	it	2				4

Longitud	d_1	d_2	d_3	d_4
Vector normal	5,068	4,899	3,762	7,738

$$d_1 = \frac{1 \times 3 + 0{,}415 \times 0{,}830}{5{,}068} = 0{,}660$$

$$d_2 = \frac{1 \times 2 + 0{,}415 \times 0}{4{,}899} = 0{,}048$$

$$d_3 = \frac{1 \times 0 + 0{,}415 \times 1{,}073}{3{,}762} = 0{,}118$$

$$d_4 = \frac{1 \times 0 + 0{,}415 \times 1{,}073}{7{,}738} = 0{,}058$$

Modelo Vectorial

- Ventajas del modelo vectorial:
  - El esquema de ponderación de términos mejora la calidad de la recuperación.
  - La estrategia de búsqueda parcial permite recuperar documentos que coinciden con los términos de búsqueda de forma parcial.
  - El ranking se calcula mediante el coseno.
  - La normalización por el tamaño del documento es inherente al modelo.
- Desventajas:
  - Se asume la independencia entre los términos.

Modelo Probabilístico

Modelo Probabilístico

- Propuesto por Robertson y Sparck Jones en 1976
- Dada una consulta, existe:
  - Un conjunto con, exactamente, los documentos relevantes
  - Otro conjunto con los documentos no relevantes
- Ambos conjuntos forman la respuesta ideal
- El proceso de consulta podría definirse como la especificación de la respuesta ideal
- ¿Cuáles son esas propiedades?
  - No lo sabemos
- Lo único que sabemos es que tenemos un conjunto de términos.
  - Emplear su semántica para obtener el resultado

Modelo Probabilístico

- Las propiedades del conjunto ideal no se conocen en el momento de formular la consulta
- Se debe realizar un esfuerzo previo averiguando cuáles son dichas propiedades
  - Se podría interactuar con el usuario
  - El usuario podría mirar los documentos recuperados
  - Decidir cuáles son relevantes y cuáles no
  - El sistema usará esta información para refinar la definición de conjunto ideal
  - Tras la repetición del proceso, deberíamos de obtener una mayor precisión

Modelo Probabilístico

- **Definición. Principio de Ranking Probabilístico**. *Dada una consulta de usuario q y un documento d_j de la colección, el modelo probabilístico trata de estimar la probabilidad de que el usuario encuentre el documento d_j interesante (relevante). El modelo asume que dicha probabilidad de relevancia depende sólo de la consulta y de la representación del documento, es decir, sólo depende de la información contenida en el documento. Además, el modelo asume que existe un subconjunto de todos los documentos que el usuario prefiere como conjunto respuesta para la consulta q. De este modo, el conjunto de la respuesta ideal se llamará R y deberá maximizar la probabilidad global de relevancia del usuario. Los documentos en el conjunto R son predichos como relevantes para la consulta. Los documentos que no están en el conjunto R, son predichos como no relevantes.*

Modelo Probabilístico

- La definición es problemática
  - La relevancia para el usuario se puede ver afectada por factores externos al sistema
- La respuesta ideal producida a partir de la información del sistema puede que no sea *ideal* desde el punto de vista del usuario
- No se especifica como se tiene que calcular la probabilidad de relevancia
- Ni siquiera se da el espacio muestral a utilizar
- Dada una consulta q, el modelo probabilístico asigna a cada documento d_j, como medida de similitud a la consulta, el ratio $P(d_j\ relevante - para\ q)/P(d_j\ no - relevante - para\ q)$

Modelo Probabilístico

Definición

- En el modelo probabilístico, una consulta q es un subconjunto de términos de indexación.
- Un documento d_j se representara como un vector de pesos binarios indicando la presencia o ausencia de los términos de indexación:

$$\vec{d} = (w_{1,j}, w_{2,j}, \dots, w_{t,j})$$

- Donde $w_{i,j} = 1$ si el término k_j aparece en el documento d_j, y $w_{i,j} = 0$ en otro caso.
- Sea R el conjunto de documentos conocidos (o averiguados inicialmente) que son relevantes para el usuario, para una consulta dada q.
- Sea $\overline{R}$el complemento de R (conjunto de documentos no relevantes).

Modelo Probabilístico

- $P(R|\vec{d}_{j,q})$ es la probabilidad de que el documento d_j con representación $\vec{d}_j$ sea relevante para la consulta q. Asimismo, $P(\bar{R}|\vec{d}_{j,q})$ es la probabilidad de que el documento $\vec{d}_j$ no sea relevante para q.
- La similitud $sim(d_j, q)$ de un documento d_j para un consulta q se define como:

$$sim(d_j, q) = \frac{P(R|\vec{d}_{j,q})}{P(\bar{R}|\vec{d}_{j,q})}$$

- Utilizando la regla de Bayes:

$$sim(d_j, q) = \frac{P(\overrightarrow{d_j}|R, q) \cdot P(R, q)}{P(\overrightarrow{d_j}|\bar{R}, q) \cdot P(R, q)} = \frac{P(\overrightarrow{d_j}|R, q) \cdot P(R|q)}{P(\overrightarrow{d_j}|\bar{R}, q) \cdot P(R|q)}$$

Modelo Probabilístico

- $P(\overrightarrow{d_j}|R,q)$ simboliza la probabilidad de que un documento seleccionado al azar del conjunto R de documentos relevantes (para la consulta q) tenga la representación $\overrightarrow{d_j}$
- $P(R,q)$ representa la probabilidad de que un documento seleccionado al azar del la colección completa sea relevante para la consulta q
- La ecuación se puede simplificar ya que $P(R,q)$ y $P(\bar{R},q)$ son los mismos para todos los documentos

$$sim(d_j,q) \sim \frac{P(\overrightarrow{d_j}|R,q)}{P(\overrightarrow{d_j}|\bar{R},q)}$$

Modelo Probabilístico

- $\overrightarrow{d_j}$ está compuesto por pesos binarios
  - Presencia o ausencia
- Si asumimos la independencia entre los términos

$$sim(d_j, q) \sim \frac{\left(\prod_{k_i|w_{i,j}=1} P(k_i|R,q)\right) \cdot \left(\prod_{k_i|w_{i,j}=0} P\left(\overline{k_i}|R,q\right)\right)}{\left(\prod_{k_i|w_{i,j}=1} P(k_i|\bar{R},q)\right) \cdot \left(\prod_{k_i|w_{i,j}=0} P\left(\overline{k_i}|\bar{R},q\right)\right)}$$

- $P(k_i|R,q)$ representa la probabilidad de que el término de indexación k_i esté presente en un documento seleccionado al azar del conjunto de documentos relevantes R.
- $P(k_i|\bar{R},q)$ representa la probabilidad de que el término de indexación k_i no esté presente en un documento seleccionado al azar del conjunto de documentos relevantes R.

Modelo Probabilístico

- Para simplificar la fórmula:

$$p_{iR} = P(k_i|R,q)$$
$$q_{iR} = P(k_i|\bar{R},q)$$

- Dado que, $P(k_i|R,q) + P\left(\bar{k}_i\middle|R,q\right) = 1$ y

$P(k_i|\bar{R},q) + P\left(\bar{k}_i\middle|\bar{R},q\right) = 1,$

$$sim(d_j,q) \sim \frac{\left(\prod_{k_i|w_{i,j}=1} p_{iR}\right) \cdot \left(\prod_{k_i|w_{i,j}=0}(1-p_{iR})\right)}{\left(\prod_{k_i|w_{i,j}=1} q_{iR}\right) \cdot \left(\prod_{k_i|w_{i,j}=0}(1-q_{iR})\right)}$$

- Tomando logaritmos (cambia el valor absoluto del ranking, no el valor de ranking)

$$sim(d_j,q) \sim \log \prod_{k_i|w_{i,j}=1} p_{iR} + \log \prod_{k_i|w_{i,j}=0} (1-p_{iR})$$
$$- \log \prod_{k_i|w_{i,j}=1} q_{iR} - \log \prod_{k_i|w_{i,j}=0} (1-q_{iR})$$

Modelo Probabilístico

- Tras un proceso de simplificación de la fórmula, obtenemos:

$$sim(d_j, q) \sim \sum_{k_i \in q \wedge k_i \in d_j} \log\left(\frac{p_{iR}}{1 - p_{ir}}\right) + \log\left(\frac{1 - q_{iR}}{q_{ir}}\right)$$

- Al principio, el conjunto R no es conocido
- Solución: diseñar un método para calcular las probabilidades iniciales p_{iR} y q_{iR}
- Existen diferentes alternativas:
  - Tabla de contingencia de incidencia de términos
  - Ranking en ausencia de información de relevancia

Modelo Probabilístico
Tabla de contingencia de incidencia de términos

- **Definición**. Sea N el número de documentos en una colección y n_i el número de documentos que contienen al término k_i. Sea R el número total de documentos relevantes para la consulta q (en opinión de los usuarios), y r_i el número de documentos relevantes que contienen al término k_i.
- La tabla de contingencia de incidencias de términos puede definirse como:

Caso	Relevantes	No relevantes	Total
Documentos conteniendo a k_i	r_i	$n_i - r_i$	n_i
Documentos no conteniendo a k_i	$R - r_i$	$N - n_i - (R - r_i)$	$N - n_i$
Todos los documentos	R	$N - R$	N

Modelo Probabilístico

- Si suponemos que la información de la tabla está disponible para todas las consultas:

$$p_{iR} = \frac{r_i}{R}, q_{iR} = \frac{n_i r_i}{N-R}$$

- La formula quedaría

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

- Para valores pequeños de r_i es conveniente añadir 0,5 a cada término

$$sim(d_j, q) \sim \sum_{k_i \in q \wedge k_i \in d_j} \log\left(\frac{(r_i + 0{,}5)(N - n_i - R + r_i + 0{,}5)}{(R - r_i + 0{,}5) - (n_i - r_i + 0{,}5)}\right)$$

Modelo Probabilístico

- Es necesario conocer los valores de r_i y R para poder resolver la fórmula
- Solución: asumir que $\mathrm{R} = r_i = 0$

$$sim(d_j, q) \sim \sum_{k_i \in q \wedge k_i \in d_j} \log\left(\frac{(N - n_i + 0{,}5)}{n_i + 0{,}5}\right)$$

- En ausencia de información sobre la relevancia, esta sería la fórmula
- El IDF está presente
- Ni TF ni la longitud normalizada del documento están presentes
- Dichos problemas se solucionan en el modelo probabilístico BM25

Modelo Probabilístico

- Ejemplo. Consulta “to do”.

documento	Cálculo del ranking	Ranking
d_1	$log\frac{4-2+0{,}5}{2+0{,}5}+log\frac{4-3+0{,}5}{3+0{,}5}$	$-1{,}222$
d_2	$log\frac{4-2+0{,}5}{2+0{,}5}$	0
d_3	$log\frac{4-3+0{,}5}{3+0{,}5}$	$-1{,}222$
d_4	$log\frac{4-3+0{,}5}{3+0{,}5}$	$-1{,}222$

Modelo Probabilístico

#	Término	IDF	d_1	d_2	d_3	d_4
1	to	2	*	*		
2	do	3	*		*	*
3	is	1	*			
4	be	0				
5	or	1		*		
6	not	1		*		
7	I	2		*	*	
8	am	2		*	*	
9	what	1		*		
10	think	1			*	
11	therefore	1			*	
12	da	1				*
13	let	1				*
14	it	1				*

$$d_1 = log\frac{4-2+0{,}5}{2+0{,}5} + log\frac{4-3+0{,}5}{3+0{,}5} = -1{,}222$$

$$d_2 = log\frac{4-2+0{,}5}{2+0{,}5} = 0$$

$$d_3 = log\frac{4-3+0{,}5}{3+0{,}5} = -1{,}222$$

$$d_4 = log\frac{4-3+0{,}5}{3+0{,}5} = -1{,}222$$

Modelo Probabilístico

- Los resultados contienen valores negativos
- La formula no funciona de forma correcta cuando $n_i > n/2$
  - En dicho caso se introducen valores negativos
- Solución:
  - Eliminar el valor n_i del numerador

$$sim(d_j, q) \sim \sum_{k_i \in q \wedge k_i \in d_j} \log\left(\frac{N + 0{,}5}{n_i + 0{,}5}\right)$$

Modelo Probabilístico

- Ejemplo. Consulta “to do”.

documento	Cálculo del ranking	Ranking
d_1	$log\frac{4+0{,}5}{2+0{,}5}+log\frac{4+0{,}5}{3+0{,}5}$	1,210
d_2	$log\frac{4+0{,}5}{2+0{,}5}$	0,847
d_3	$log\frac{4+0{,}5}{3+0{,}5}$	0,362
d_4	$log\frac{4+0{,}5}{3+0{,}5}$	0,362

- * los documentos 3 y 4 tienen el mismo peso debido a que no se normaliza por el tamaño del documento

Modelo Probabilístico

- Ventajas: en teoría el modelo es optimo y devuelve los documentos ordenados por probabilidad de ser relevantes
- Desventajas
  - La necesidad de estimar previamente la separación entre documentos relevantes y no relevantes
  - No se tiene en cuenta la frecuencia de aparición de un término dentro de un documento
  - No se normaliza por el tamaño del documento

Modelos Alternativos

Modelos Alternativos: Conjuntos

Modelos Alternativos: conjuntos

- Destacamos tres modelos alternativos que se basan en la teoría de conjuntos
  1. Modelos basados en conjuntos
  - Combina teoría de conjuntos con el ranking del espacio vectorial
  - Aunque tiene características booleanas y algebraicas, lo consideraremos como booleano
  - Se basa en las dependencias mutuas entre los términos de indexación para mejorar los resultados
  - Las dependencias se basan en la correlación entre términos

Modelos Alternativos: conjuntos

2. Modelo booleano extendido
   - Búsqueda de correspondencias parcial y ponderación de términos
   - Combina el modelo booleano con el modelo vectorial
3. Modelo basado en conjuntos difusos
   - La representación de documentos y consultas a través de un conjunto de palabras da como resultado descripciones que sólo están parcialmente relacionadas con el contenido real.
   - La relación entre un documento y la consulta es aproximada
   - Solución: cada término de la consulta define un conjunto difuso, y por tanto cada documento tiene un grado de pertenencia a dicho conjunto

Modelos Alternativos: Algebraicos

Modelos Alternativos: algebraicos

- Destacamos tres modelos alternativos algebraicos:
  1. Espacio vectorial generalizado
  - El modelo vectorial clásico asume la independencia de los términos
  - La independencia entre los términos de indexación suele entenderse como la ortogonalidad entre vectores
  - Este modelo se basa en que los vectores están compuestos por pequeños componentes derivados de una colección particular

Modelos Alternativos: algebraicos

2. Semántica latente

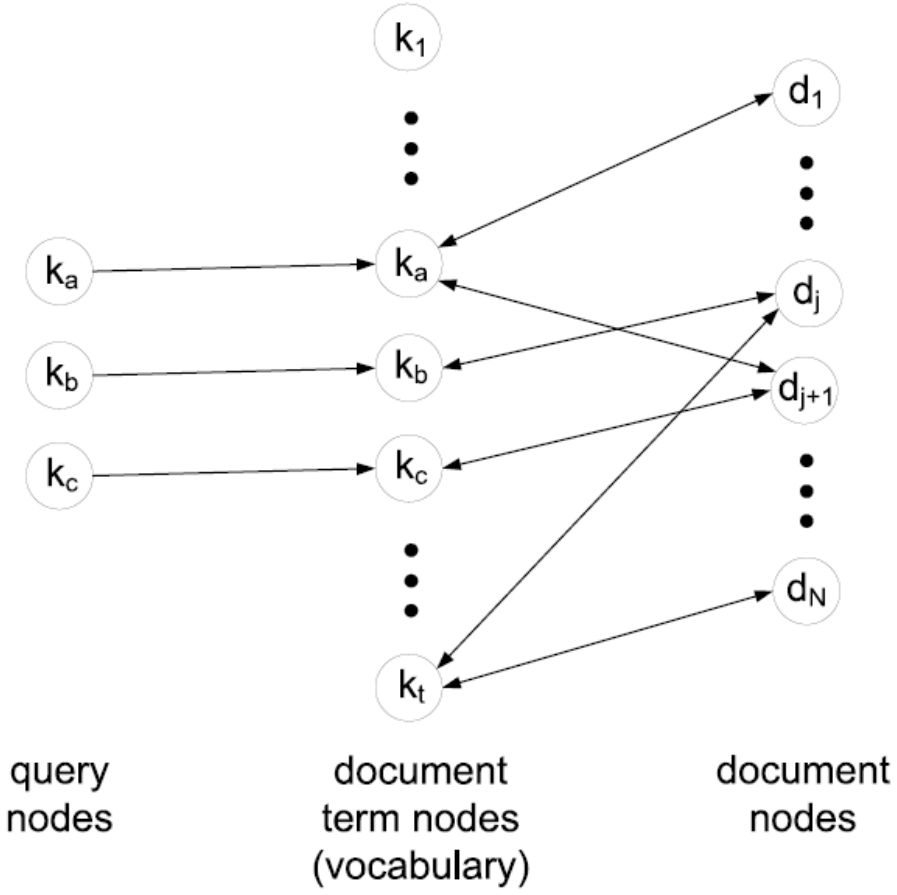
- Representar los documentos y consultas como términos puede provocar una recuperación de baja calidad
  - Documentos sin relación
  - Documentos relacionados pero que no están indexados por los términos de la consulta no serán devueltos
- El texto de un documento contiene, conceptos y relaciones entre ellos
- Se basa en la asociación del vector de cada documento y consulta a un espacio dimensional compuesto por conceptos

Modelos Alternativos: algebraicos

3. Redes neuronales

- El cerebro humano se compone de millones de neuronas
- Cada neurona es una unidad de procesamiento única
  - Se estimulan con una entrada y devuelven una salida
- La salida de una neurona alimenta la entrada de otra neurona
- Este proceso se repite a través de una capa de neuronas (proceso de propagación de la activación)
- Como resultado, la información de entrada es procesada lo que deriva en una reacción física como respuesta
- Una red neuronal es una simplificación del grafo de interconexiones entre las neuronas del cerebro humano

Modelos Alternativos: algebraicos



Modelos Alternativos: Probabilísticos

Modelos Alternativos: probabilísticos

- Destacaremos tres principales modelos probabilísticos
  1. BM25
  - Extensión del modelo clásico para tener en cuenta la ponderación de los términos y la longitud del documento
  2. Modelos del lenguaje
  - Definen la probabilidades de distribución por documentos, usándolas para predecir probabilidad de observar los términos de la consulta
  3. Divergencia con respecto a la aleatoriedad
  - Se basa en la medición de la divergencia entre la distribución de un término producida por un proceso al azar y la distribución real del término