Método de ranking en el diseño de un sistema de acceso a la información

Doble Grado en Ingeniería Informática y Matemáticas

Johanna Capote Robayna

8 de Julio de 2020

Trabajo Fin de Grado

E.T.S. de Ingenierías Informática y de Telecomunicación
Facultad de Ciencias

Índice

Introducción

Contextualización

- En la década de los 80 Internet empieza a darse a conocer con una primera páginas web.
- WebCrawler, Lycos, AltaVista o Yahoo fueron los primeros buscadores.
- En 1998 Sergey Brin y Lawrence Page diseñan el algoritmo PageRank.

Contextualización

- En la década de los 80 Internet empieza a darse a conocer con una primera páginas web.
- WebCrawler, Lycos, AltaVista o Yahoo fueron los primeros buscadores.
- En 1998 Sergey Brin y Lawrence Page diseñan el algoritmo PageRank.

Contextualización

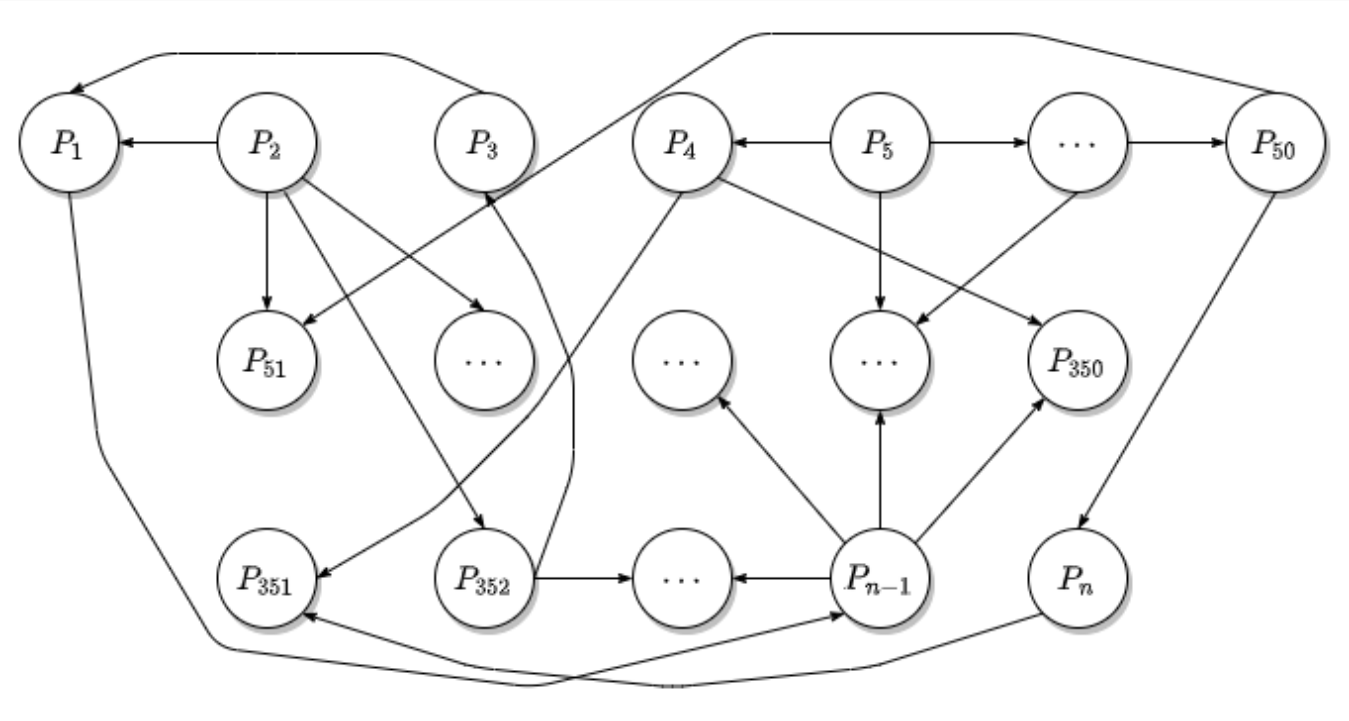
- En la década de los 80 Internet empieza a darse a conocer con una primera páginas web.
- WebCrawler, Lycos, AltaVista o Yahoo fueron los primeros buscadores.
- En 1998 Sergey Brin y Lawrence Page diseñan el algoritmo PageRank.

Matemáticas

Modelo matemático

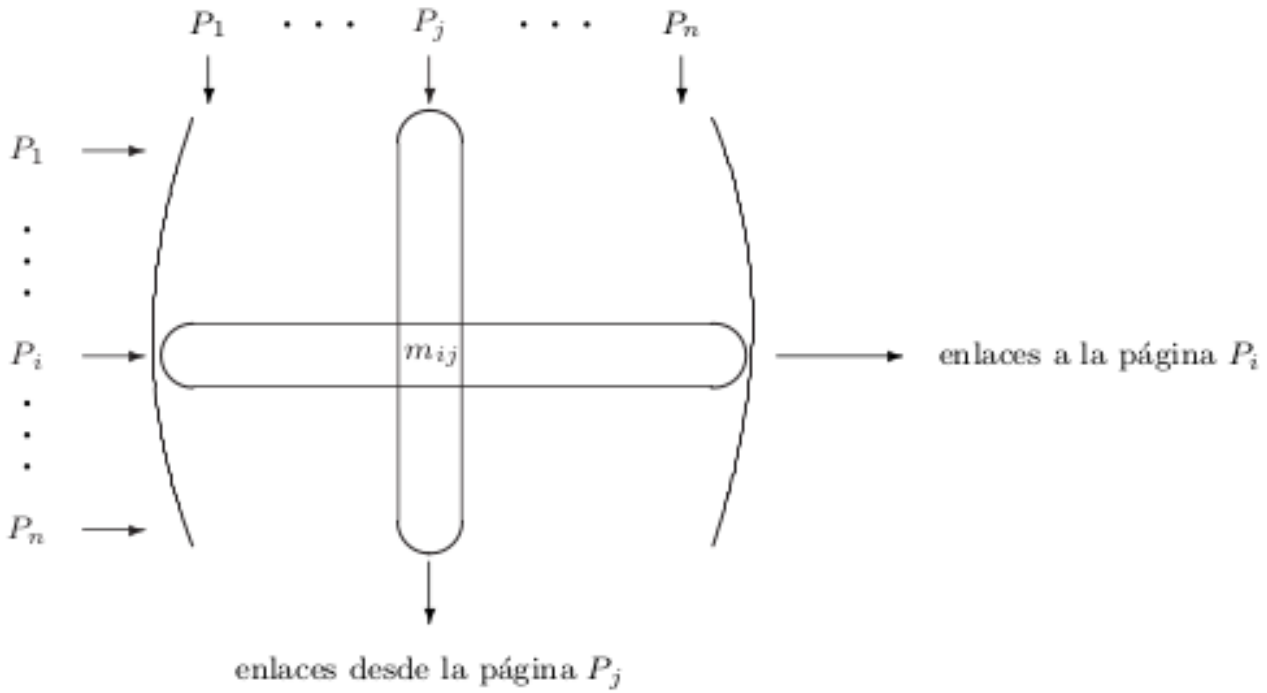
Llamamos $P_1, \dots, P_n$ a cada una de las páginas, $n \in \mathbb{N}$. Definimos la importancia $x_i \in [0,1] \subseteq \mathbb{R}$ de la página P_i.

Construimos un grafo dirigido donde representamos los enlaces entre las páginas web. Representamos cada página P_i como un nodo y cada enlace de P_i a P_j añadimos una arista de P_i a P_j con una punta de flecha.



Modelo matemático

Podemos reflejar esta información en una matriz M. Tanto en las filas como en las columnas representamos las n páginas y por cada enlace entre una página j a otra página i, escribimos un 1 en la entrada de la matriz m_{ij} y en el caso de que no haya enlace escribimos un 0.



Modelo matemático

- Primera aproximación: la importancia de una página web es proporcional al número de enlaces entrantes.
- Problema: el número de enlaces no representa del todo la importancia. No es lo mismo que la página este citada por una página cualquiera que por una página "importante" como **www.facebook.com** o **www.apple.com**.
- Solución: definimos que la importancia de una página web x_j es proporcional a la suma de las importancias de las páginas que enlazan con P_j.

Supongamos, por ejemplo, que la página P_1 es citada desde las páginas P_{200} y P_n, que P_2 se cita desde P_1, P_{200} y P_{n-1}, mientras que en la última página P_n hay enlaces desde P_1, P_2, P_{50}, P_{200} y P_{n-1}. En nuestra asignación anterior, $x_1, \ldots, x_n$ deberían cumplir entonces que:

$$x_1 = K(x_{200} + x_n)$$

$$x_2 = K(x_1 + x_{200} + x_{n-1})$$

$$\vdots$$

$$x_n = K(x_1 + x_2 + x_{50} + x_{200} + x_{n-1})$$

donde K es una constante de proporcionalidad.

Escribimos el sistema de ecuaciones anterior en términos matriciales.

$$\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = K \begin{pmatrix} \begin{matrix} P_1 & P_2 \\ \downarrow & \downarrow \end{matrix} & & & \begin{matrix} P_{50} \\ \downarrow \end{matrix} & & & \begin{matrix} P_{200} \\ \downarrow \end{matrix} & & \begin{matrix} P_{n-1} & P_n \\ \downarrow & \downarrow \end{matrix} \\ 0 & 0 & \dots & 0 & \dots & \dots & 1 & \dots & 0 & 1 \\ 1 & 0 & \dots & 0 & \dots & \dots & 1 & \dots & 1 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & \dots & 1 & \dots & \dots & 1 & \dots & 1 & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$$

Si llamamos $\vec{x}$ al vector de importancias $(x_1, \dots, x_n)$, $\lambda = \frac{1}{K}$ y M a la matriz asociada al grafo. Nos encontramos con un problema de valores propios y vectores propios:

$$M\vec{x} = \lambda\vec{x}$$

Definición (Valor propio dominante)

Sea $A \in M_n(\mathbb{C})$, se dice que tiene valor propio dominante si el espectro

$$\sigma(A) = \{\lambda_1, \lambda_2 \dots \lambda_n\} \, (r \leq n)$$

cumple:

- $\lambda_1 > 0$.
- $m(\lambda_1) = 1$.
- $|\lambda_i| < \lambda_1$ *para* $i = 2, \dots, r$.

Al valor propio dominante lo notaremos como λ_p.

Proposición

Sea $A \in M_n(\mathbb{C})$ una matriz con valor propio dominante λ_p. Entonces el límite $\lim_{k\to\infty} \frac{1}{\lambda_p^k} A^k = Q$ donde $Q \in M_n(\mathbb{C})$ que cumple:

$$ImQ = Ker(A - \lambda_p I)\,, Q^2 = Q\,, QA = AQ$$

Se dice que Q es una $proyección\ espectral\ de\ A$.

Teorema (Perron, 1907)

Sea $A \in M_n(\mathbb{C})$, $A > 0$ *con* $\lambda_p = \rho(A)$. *Entonces:*

1. $\lambda_p > 0$.
2. $\lambda_p \in \sigma(A)$ (λ_p *es llamado raíz de Perron).*
3. $m(\lambda_p) = 1$.
4. *Existe un vector propio* $\vec{v} > 0$ *tal que* $A\vec{v} = \lambda_p \vec{v}$.
5. *El* $vector\ de\ Perron$ *es el único definido como*

$$A\vec{p} = \lambda_p \vec{p}\ , \vec{p} > 0\ , y\ \|\vec{p}\|_1 = 1$$

   y, excepto los múltiplos positivos de $\vec{p}$, *no hay otros vectores propios no negativos para* A, *independientemente del valor propio.*

6. λ_p *es el único valor propio en el circulo espectral de* A.

Definición (Matriz reducible)

Se dice que una matriz $A \in M_n(\mathbb{R})$ no negativa ($A \geq 0$) es reducible si existe una permutación simétrica (de filas y columnas) que transforma A en una matriz del tipo:

$$\left(\begin{array}{c|c} A_{11} & A_{12} \\ \hline 0 & A_{22} \end{array}\right)$$

donde A_{11} y A_{22} son matrices cuadradas. Una matriz se dice irreducible cuando no es reducible.

Proposición

Sea $A \in M_n(\mathbb{R})$ una matriz no negativa ($A \leq 0$). Son equivalentes:

1. *A es irreducible.*
2. *La matriz $(I + A)^{n-1}$ es positiva.*
3. *Si A es la matriz de adyacencia de un grafo, entonces el grafo está fuertemente conectado.*

Definición (Grafo fuertemente conectado)

Un grafo dirigido G se dice fuertemente conectado si para cada par de nodos distintos P_i, P_j en G hay un camino de longitud finita que comienza en P_i y termina en P_j.

Enunciado del teorema

Teorema (Frobenius, 1908-1912)

Sea A una matriz cuadrada ($A \in M_n(\mathbb{C})$) no negativa ($A \geq 0$). Si A es irreducible, entonces para $r = \rho(A)$ se cumple que:

1. $r > 0$, $r \in \sigma(A)$ *y*
2. $m(r) = 1$.
3. *Existe un vector propio* $\vec{x} > 0$ *asociado a* r.
4. *El único vector definido por:*

$$A\vec{p} = r\vec{p} \ , \ \vec{p} > 0 \ , y \ \|\vec{p}\|_1 = 1$$

es el vector de Perron. No hay ningún vector propio no negativo para A excepto los múltiplos positivos de $\vec{p}$, independientemente del valor propio.

Método de las potencias

El método de las potencias comienza con un vector inicial $\vec{x}^{(0)} \in \mathbb{R}^n$, $\vec{x}^{(0)} \geq 0$, para calcular los siguiente $\vec{x}^{(k)}$ se utiliza la siguiente fórmula recurrente:

$$\vec{x}^{(k+1)} = A\vec{x}^{(k)} \text{ con } k \in \mathbb{N}$$

esta fórmula la podemos desarrollar obteniendo

$$\vec{x}^{(k)} = A^k\vec{x}^{(0)}$$

Se puede demostrar que

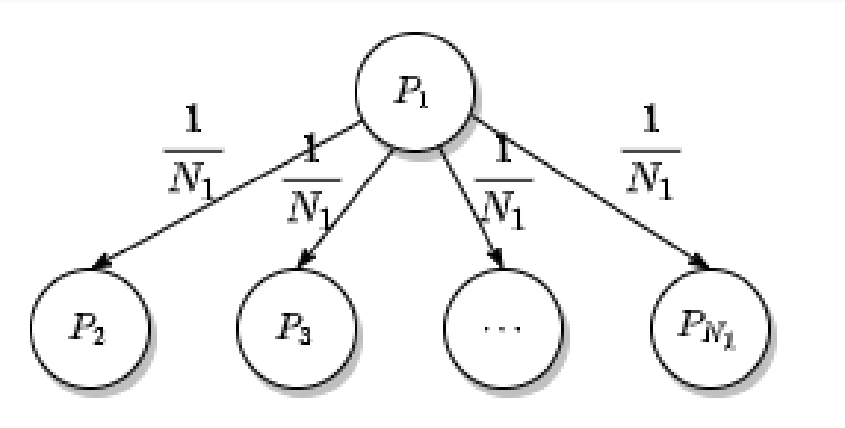
$$\lim_{k\to\infty} \frac{1}{\|\vec{x}^{(k)}\|_1}\vec{x}^{(k)} = \vec{p}$$

Siendo $\vec{p}$ el vector de Perron de A.

Informática

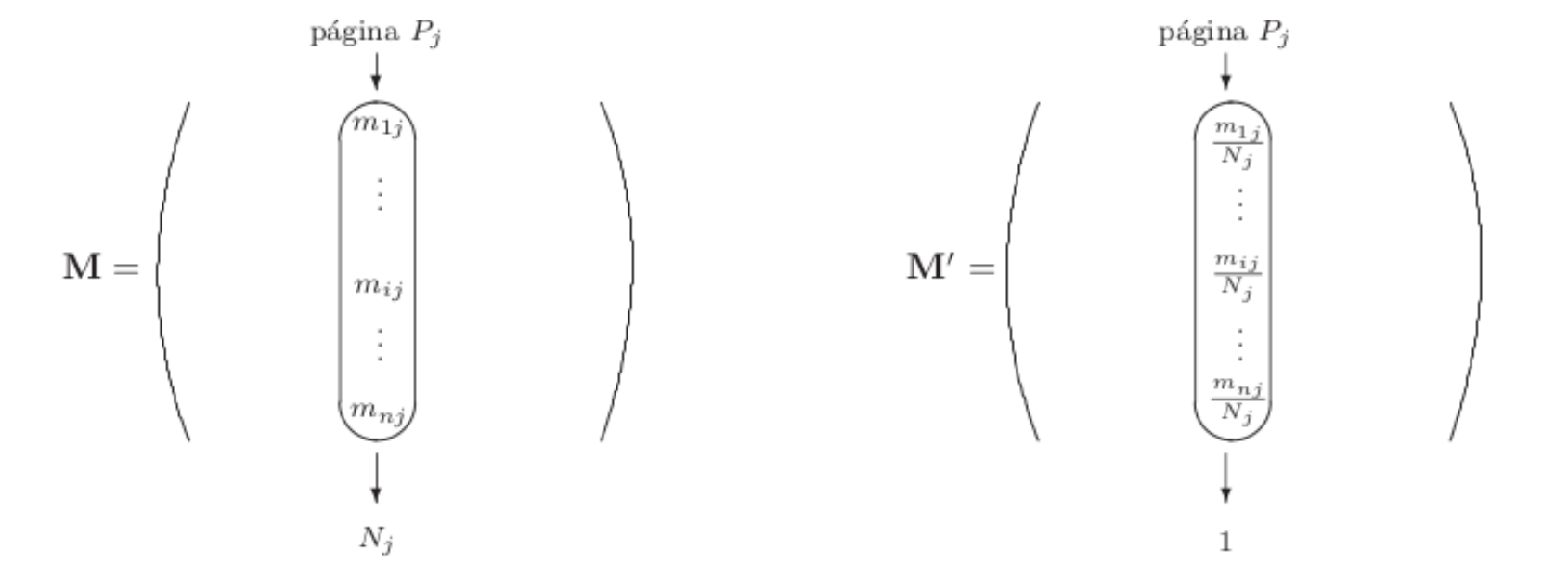
Modificaciones previas

Supongamos que el usuario se encuentra navegando por la red, por ejemplo, en la primera página P_1 y cuando se aburre decide saltar a una de las páginas con las que enlaza P_1 . Supongamos que hay N_1 páginas que enlazan desde P_1 , es lógico pensar que todas tienen una probabilidad de ser elegidas del $\frac{1}{N_1}$ es decir que siguen una distribución de probabilidad uniforme (discreta) en $[1, N_1]$.



Volvemos a considerar las mismas paginas web $P_1, P_2, \dots, P_n$ y M la matriz de adyacencia del grafo, cuyas entradas m_{ij} son 0 y 1. Llamamos N_j al número de enlaces de la página P_j , es decir al número de entradas de la columna j. Construimos una nueva matriz M' a partir de la M original sustituyendo cada m_{ij} por:

$$m'_{ij} = \frac{m_{ij}}{N_j}$$



Con este modelo podemos conocer con que probabilidad estará el usuario en cada una de las páginas tras cada instante de tiempo, entendiendo instante de tiempo como transiciones o saltos. Si multiplicamos M' por el vector inicial, obtenemos:

$$\begin{pmatrix} \cdots & \cdots & m'_{1k} & \cdots \\ \vdots & \ddots & \vdots & \vdots \\ \cdots & & m'_{kk} & \cdots \\ \vdots & \vdots & \vdots & \ddots \\ \cdots & \cdots & m'_{nk} & \cdots \end{pmatrix} \begin{pmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{pmatrix} = \begin{pmatrix} m'_{1k} \\ \vdots \\ m'_{kk} \\ \vdots \\ m'_{nk} \end{pmatrix}$$

Este vector resultante, cuyas entradas son o 0 o $\frac{1}{N_k}$, describe con que probabilidad estará el usuario en cada una de las páginas tras una unidad de tiempo.

- **Problema**: Sin embargo, podría ocurrir que alguna de las páginas no citaran a ninguna otra, es decir, que ese nodo del grafo no tuviera enlaces salientes. Esto se traduce en que en nuestra matriz M' aparece una columna de ceros, por lo que esta matriz dejaría de ser primitiva y además el grafo del que partimos no estaría fuertemente conectado.

- **Solución**:

$$M'' = cM' + (1-c) \begin{pmatrix} p_1 \\ \vdots \\ p_n \end{pmatrix} (1, \dots, 1)$$

  Donde $p_1, \dots, p_n$ es una distribución de probabilidad y c es un parámetro entre 0 y 1. La distribución de probabilidad a que elegimos es una distribución uniforme que asigne una probabilidad de $p_i = 1/n$.

Definimos:

- El conjunto de todos los términos de la colección de documentos como $V = \{t_1, t_2, \dots, t_M\}$
- El conjunto de todos los documentos de la colección como $D = \{d_1, d_2, \dots, d_N\}$.
- Para cada término t_i se define T_i como el conjunto formado por todos los documentos que contienen el término t_i.

Si por ejemplo tuvieramos los siguientes cuatro documentos:

- $d_1 =$ {gato, gato, gato, tortuga, pez}
- $d_2 =$ {perro, caballo}
- $d_3 =$ {gato, perro, águila}
- $d_4 =$ {pez, tortuga, tortuga}

Entonces la consulta obtendría el siguiente resultado:

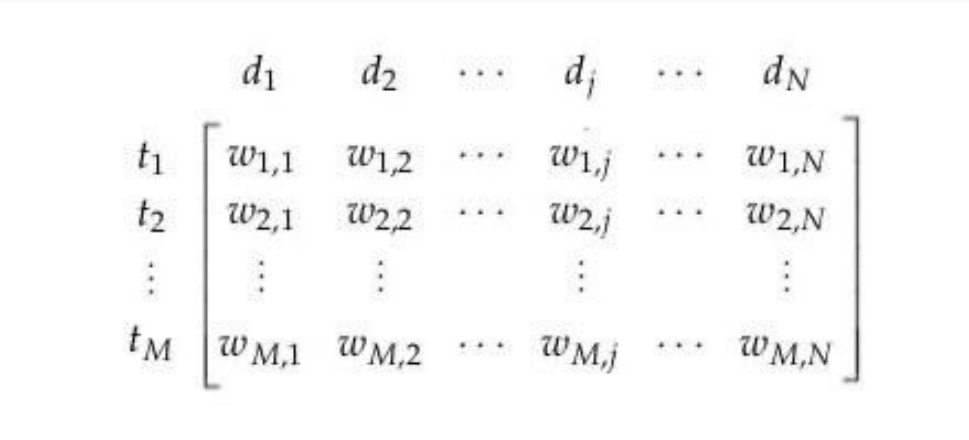
- $\text{perro AND gato} = T_1 \cap T_2 = \{d_2, d_3\} \cap \{d_1, d_3\} = \{d_3\}$
- $\text{perro OR gato} = T_1 \cup T_2 = \{d_2, d_3\} \cup \{d_1, d_3\} = \{d_1, d_2, d_3\}$
- $\text{perro OR NOT gato} = T_1 \cup (D - T_2) = \{d_2, d_3\} \cup \{d_2, d_4\} = \{d_2, d_3, d_4\}$

Modelo Booleano

- Ventajas
  - Sencillez.
  - Flexibilidad.
- Inconvenientes
  - No se tiene en cuenta la frecuencia del término en el documento.
  - No se tiene en cuenta la relevancia del término.

Definimos:

- El conjunto de todos los términos de la colección de documentos como $V = \{t_1, t_2, \dots, t_M\}$
- El conjunto de todos los documentos de la colección como $D = \{d_1, d_2, \dots, d_N\}$.
- Un documento $\vec{d}_j$ como un vector, de tamaño el número de términos, $\vec{d}_j = (w_{1,j}, w_{2,j}, \dots, w_{M,j})$ donde $w_{i,j}$ indice el peso del término i en el documento j.



Este modelo se basa en dos conceptos fundamentales:

- Esquema de pesos.
- Similitud entre dos vectores de términos (por ejemplo, entre un documento y una consulta).

$$sim(\vec{q}, \vec{d}) = cos(\alpha) = \frac{\sum_{i=1}^{M} w_{i,q} \cdot w_{i,d}}{\sqrt{\sum_{i=1}^{M} w_{i,d}^2} \cdot \sqrt{\sum_{i=1}^{M} w_{i,q}^2}} = \frac{\vec{q}}{|\vec{q}|} \cdot \frac{\vec{d}}{|\vec{d}|}$$

$$\text{donde } |\vec{v}| = \sqrt{\sum_{i=1}^{M} v_i^2}$$

A la hora de calcular el peso de un término en cada documento debemos tener en cuenta dos conceptos:

- La frecuencia del término en el documento.
- La especificidad del término en el documento.

Por lo tanto, calcularemos el peso de un término de la siguiente forma:

$$w_{i,j} = tf_{i,j} \cdot idf_i = tf_{i,j} \cdot \log\left(\frac{N}{n_i}\right)$$

Tras calcular estos pesos se normaliza cada vector $\vec{d}_j$:

$$\vec{d}_j = \frac{\vec{d}_j}{|\vec{d}_j|}$$

Procedimiento:

- **Antes de introducir la consulta:**
  - Se calculan la matriz de pesos.
- **Introducida la consulta:**
  - Se calculan los pesos de la consulta $\vec{q}$.
  - Se calcula la similitud de la consulta $\vec{q}$ con cada documento de la colección. Se obtiene un vector de relevancias donde queda reflejada la similitud de cada documentos a la consulta.
  - Se multiplica el vector de relevancias por el vector del PageRank.

Realimentación de consultas

El objetivo de esta técnica consiste en conseguir una consulta expandida $\vec{q}'$ a partir de una consulta $\vec{q}$, para ello se utiliza la fórmula de Rocchio:

$$\vec{q}' = \alpha\vec{q} + \frac{\beta}{D_r}\sum_{\vec{d}_j \in D_r}\vec{d}_j - \frac{\gamma}{D_{nr}}\sum_{\vec{d}_j \in D_{nr}}\vec{d}_j$$

Usualmente se le da valores $\alpha = 1$, $\beta = 0.75$ y $\gamma = 0$, ya que no buscamos ampliar la consulta con documentos no relevantes.

Ejemplo

Quinto clasificado

- Título: Neuroimaging Endpoints in Amyotrophic Lateral Sclerosis.
- ORCID autor: 0000-0003-0267-3180
- Abstract: Amyotrophic lateral sclerosis ...

Decimo clasificado

- Título: Pathology and MRI: exploring cognitive impairment in MS.
- ORCIRD autor: 0000-0002-6378-0070
- Abstract: Cognitive impairment is a frequent symptom in people with multiple sclerosis, affecting up to 70 % of patients ...

Conclusiones y trabajos futuros

- Conclusiones:
  - Se resuelve el problema de ordenar información según el interés del usuario.
  - Se consiguen personalizar las búsquedas.
- Trabajos futuros:
  - Crear perfiles de usuario.
  - Modificar el software para que sea más escalable.
  - Paralelizar el software para conseguir trabajar con conjunto más grandes.

Gracias por su atención