

Ponderação de termos

Wendel Melo

Faculdade de Computação
Universidade Federal de Uberlândia

Recuperação da Informação

Adaptado do Material da Profª Vanessa Braganholo - IC/UFF

Ponderação de Termos

- Parte da ideia de que, dentro de um contexto, alguns termos podem ser mais importantes do que outros para descrever o conteúdo dos documentos;
- Por exemplo, um termo que apareça em todos os documentos da base não tem muita utilidade para indexação. Por outro lado, um termo raro pode ter grande importância;
- Por sua vez, um termo que apareça muitas vezes em um documento específico pode, em muitos casos, dar uma ideia melhor sobre o conteúdo desse documento do que um termo que apareça poucas vezes.

Ponderação de Termos

- Assim, para cada termo k_i em um documento d_j , pode-se associar um peso numérico $w_{ij} \geq 0$ que quantifica a importância de k_i na descrição do conteúdo de d_j ;
- Observe então que um mesmo termo pode ter diferentes pesos em diferentes documentos;
- Assim, w será uma matriz de pesos onde, em cada coluna j , temos o vetor de pesos do documento d_j : $(w_{1j}, w_{2j}, \dots, w_{Tj})$.

Ponderação de Termos

- A ponderação de termos visa atribuir, em cada documento, um peso numérico para cada um dos termos do vocabulário;
- Idealmente, o peso de um termo k_i em um documento d_j deve ser proporcional à *importância* de k_i em d_j (o conceito de “importância” pode variar, todavia);

Ponderação de Termos

- **A ponderação de termos, por si só, não tem por objetivo ranquear documentos**, apenas quantificar a importância de cada termo em cada documento!
- A ponderação de termos nos documentos é realizada logo após a etapa de criação do índice invertido! É feita **antes** de um sistema de RI entrar em operação!

Ponderação de Termos

Definimos:

- N : número de documentos na base;
- f_{ij} : frequência de ocorrência do termo k_i no documento d_j (quantidade de vezes em que k_i aparece em d_j);
- F_i : frequência total do termo k_i em toda a base de documentos;
- n_i : número de documentos onde o termo k_i aparece ao menos uma vez.

Ponderação de Termos

- Partindo das medidas anteriores, define-se três formas clássicas de ponderação de termos:
 - TF (*Term Frequency* - Frequência de Termo);
 - IDF (*Inverse Document Frequency* - Frequência Inversa de Documento);
 - TF-IDF.

Ponderação TF

- TF (*Term Frequency* – Frequência de Termo) se baseia na premissa de que quanto mais vezes um termo aparece em um documento, maior sua capacidade de descrever seu conteúdo.
- Assim, o peso w_{ij} do termo k_i no documento d_j é proporcional a frequência f_{ij} :

$$w_{ij} \sim f_{ij}$$

Ponderação TF

- Podemos então usar a ponderação TF para calcular os pesos w_{ij} (peso do termo k_i no doc d_j) da seguinte forma:

$$w_{ij} = \begin{cases} 1 + \log f_{ij} & , \text{ se } f_{ij} \geq 1 \\ 0 & , \text{ caso contrário} \end{cases}$$

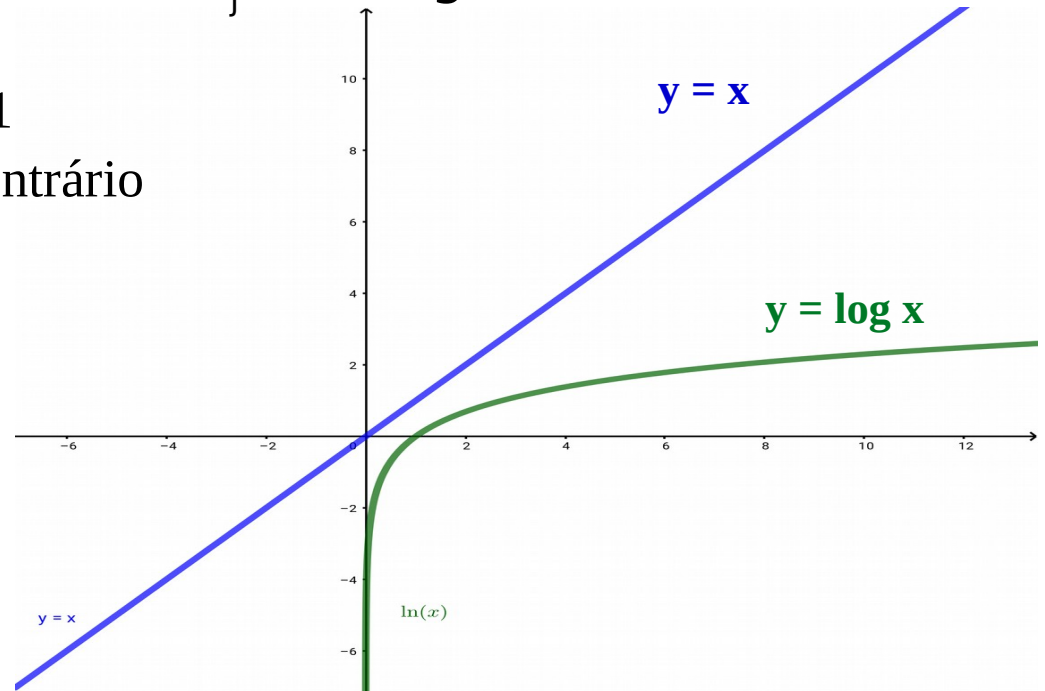
Onde f_{ij} é a frequência do termo k_i no documento d_j .

Ponderação TF

- Podemos então usar a ponderação TF para calcular os pesos w_{ij} (peso do termo k_i no doc d_j) da seguinte forma:

$$w_{ij} = \begin{cases} 1 + \log f_{ij} & , \text{ se } f_{ij} \geq 1 \\ 0 & , \text{ caso contrário} \end{cases}$$

Observe que o uso do logaritmo torna o crescimento da função menos acentuado



Ponderação TF

- Uma variante da ponderação TF utiliza a frequência normalizada:

$$w_{ij} = \frac{f_{ij}}{\max_p f_{pj}}$$

Nessa variante, divide-se cada frequência pela maior frequência de termo do documento.

Ponderação IDF

- IDF (*Inverse Document Frequency* – Frequência Inversa de Documento) busca expressar a importância de um termo k_i dentro da base de documentos segundo sua raridade.

$$idf(k_i) = \log\left(\frac{N}{n_i}\right)$$

onde:

- N : número total de documentos
- n_i : número de documentos com o termo k_i

Ponderação IDF

- IDF (*Inverse Document Frequency* – Frequência Inversa de Documento) busca expressar a importância de um termo k_i dentro da base de documentos segundo sua raridade.

$$idf(k_i) = \log\left(\frac{N}{n_i}\right)$$

Quanto mais raro é um termo, maior o seu idf!

onde:

- N : número total de documentos
- n_i : número de documentos com o termo k_i

Ponderação IDF

- IDF (*Inverse Document Frequency* – Frequência Inversa de Documento) busca expressar a importância de um termo k_i dentro da base de documentos segundo sua raridade.

$$idf(k_i) = \log\left(\frac{N}{n_i}\right)$$

Quanto mais raro é um termo, maior o seu idf!

onde:

- N : número total de documentos
- n_i : número de documentos com o termo k_i
- *O valor idf de um termo não varia conforme o documento!*

Ponderação IDF

- Podemos então usar a ponderação IDF para calcular os pesos w_{ij} (peso do termo k_i no doc d_j) da seguinte forma:

$$w_{ij} = \begin{cases} \text{idf}(k_i) & , \text{ se } f_{ij} \geq 1 \\ 0 & , \text{ caso contrário} \end{cases}$$

Onde f_{ij} é a frequência do termo k_i no documento d_j .

Ponderação TF-IDF

- Ponderação TF: privilegia os termos que mais aparecem em cada documento;
- Ponderação IDF: privilegia os termos mais raros na base como um todo;
- Qual dos esquemas de ponderação parece mais apropriado?

Ponderação TF-IDF

- Ponderação TF: privilegia os termos que mais aparecem em cada documento;
- Ponderação IDF: privilegia os termos mais raros na base como um todo;
- Qual dos esquemas de ponderação parece mais apropriado?
 - **Talvez ambos!**

Ponderação TF-IDF

- A Ponderação TF-IDF mescla TF e IDF num só esquema para premiar a frequência dos termos no documento em conjunto com sua raridade na base;
- Assim, os pesos w_{ij} (peso do termo k_i no doc d_j) são calculados da seguinte forma:

$$w_{ij} = \begin{cases} \text{tf}(k_i, d_j) \times \text{idf}(k_i) & , \text{ se } f_{ij} \geq 1 \\ 0 & , \text{ caso contrário} \end{cases}$$

Ponderação TF-IDF

- A Ponderação TF-IDF mescla TF e IDF num só esquema para premiar a frequência dos termos no documento em conjunto com sua raridade na base;
- Assim, os pesos w_{ij} (peso do termo k_i no doc d_j) são calculados da seguinte forma:

$$w_{ij} = \begin{cases} \text{tf}(k_i, d_j) \times \text{idf}(k_i) & , \text{ se } f_{ij} \geq 1 \\ 0 & , \text{ caso contrário} \end{cases}$$

ou:

$$w_{ij} = \begin{cases} (1 + \log(f_{ij})) \times \log\left(\frac{N}{n_i}\right) & , \text{ se } f_{ij} \geq 1 \\ 0 & , \text{ caso contrário} \end{cases}$$

N : número de docs ;
 f_{ij} : frequência do termo k_i no doc d_j ;
 n_i : número de docs com o termo k_i .

Ponderação TF-IDF

- TF-IDF é o esquema de ponderação mais popular na prática;
- Em geral, termos com TF alto em um documento tendem a apresentar IDF baixo, e vice-versa (podem haver exceções);
- Assim, um alto TF pode ser equilibrado com um baixo IDF, e vice-versa;
- Comumente, os termos de maior TF-IDF são termos com valores intermediários de IDF que aparecem muitas vezes em um documento.

Exemplo - Ponderação TF-IDF

- Considere uma base de 4 documentos com os termos A, B e C:

Documento	Conteúdo
1	A A A B
2	A A C
3	A A
4	B B

Exemplo - Ponderação TF-IDF

- Considere uma base de 4 documentos com os termos A, B e C:

Documento	Conteúdo
1	A A A B
2	A A C
3	A A
4	B B

Começamos pelo cálculo dos
IDFs, pois não variam de
acordo com o documento

Exemplo - Ponderação TF-IDF

- Considere uma base de 4 documentos com os termos A, B e C:

Documento	Conteúdo
1	A A A B
2	A A C
3	A A
4	B B

$$idf(A) = \log\left(\frac{N}{n_t}\right) = \log\left(\frac{4}{3}\right) = 0.1249$$

$$idf(B) = \log\left(\frac{4}{2}\right) = 0.3010$$

$$idf(C) = \log\left(\frac{4}{1}\right) = 0.6021$$

N : número de docs ;

f_{ij} : frequência do termo k_i no doc d_j ;

n_i : número de docs com o termo k_i .

Exemplo - Ponderação TF-IDF

- Considere uma base de 4 documentos com os termos A, B e C:

Documento	Conteúdo
1	A A A B
2	A A C
3	A A
4	B B

Calculamos os pesos no documentos segundo a expressão TF-IDF:

$$w_{ij} = \begin{cases} (1 + \log(f_{ij})) \times idf(k_i) & , \text{ se } f_{ij} \geq 1 \\ 0 & , \text{ caso contrário} \end{cases}$$

N : número de docs ;

f_{ij} : frequência do termo k_i no doc d_j ;

n_i : número de docs com o termo k_i .

Exemplo - Ponderação TF-IDF

- Considere uma base de 4 documentos com os termos A, B e C:

Documento	Conteúdo
1	A A A B
2	A A C
3	A A
4	B B

$$w_{A1} = (1 + \log 3) * 0.1249 = 0.1845$$

$$w_{B1} = (1 + \log 1) * 0.3010 = 0.3010$$

$$w_{C1} = 0$$

Vetor de pesos do
documento 1

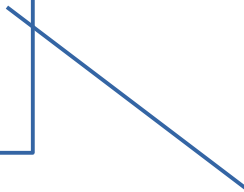
$$\bar{D}_1 = (0.1845, 0.3010, 0)$$

Exemplo - Ponderação TF-IDF

- Considere uma base de 4 documentos com os termos A, B e C:

Documento	Conteúdo
1	A A A B
2	A A C
3	A A
4	B B

Vetor de pesos do
documento 2



$$w_{A1} = (1 + \log 3) * 0.1249 = 0.1845$$

$$w_{B1} = (1 + \log 1) * 0.3010 = 0.3010$$

$$w_{C1} = 0$$

$$\bar{D}_1 = (0.1845, 0.3010, 0)$$

$$w_{A2} = (1 + \log 2) * 0.1249 = 0.1625$$

$$w_{B2} = 0$$

$$w_{C2} = (1 + \log 1) * 0.6021 = 0.6021$$

$$\bar{D}_2 = (0.1625, 0, 0.6021)$$

Exemplo - Ponderação TF-IDF

- Considere uma base de 4 documentos com os termos A, B e C:

Documento	Conteúdo
1	A A A B
2	A A C
3	A A
4	B B

$$w_{A3} = (1 + \log 2) * 0.1249 = 0.1625$$

$$w_{B3} = 0$$

$$w_{C3} = 0$$

Vetor de pesos do
documento 3

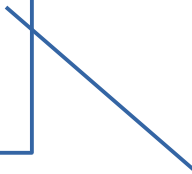
$$\bar{D}_3 = (0.1625, 0, 0)$$

Exemplo - Ponderação TF-IDF

- Considere uma base de 4 documentos com os termos A, B e C:

Documento	Conteúdo
1	A A A B
2	A A C
3	A A
4	B B

Vetor de pesos do
documento 4



$$w_{A3} = (1 + \log 2) * 0.1249 = 0.1625$$

$$w_{B3} = 0$$

$$w_{C3} = 0$$

$$\bar{D}_3 = (0.1625, 0, 0)$$

$$w_{A4} = 0$$

$$w_{B4} = (1 + \log 2) * 0.3010 = 0.3916$$

$$w_{C4} = 0$$

$$\bar{D}_4 = (0, 0.3916, 0)$$

Ponderação TF-IDF

- O uso de ponderação tende a melhorar a qualidade de um sistema de RI;
- Na prática, também é preciso adotar um vetor de pesos para a consulta. A mesma ponderação dos documentos pode ser utilizada;
- Os esquemas considerados aqui consideram os termos como totalmente independentes. No mundo real, pode ser vantajoso considerar correlação entre os diferentes termos em alguns contextos.